Expert Systems With Applications 280 (2025) 127077

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

# Multi-modal wound classification using wound image and location by Swin Transformer and Transformer

Ramin Mousa [a],*, Behnaz Rezaei [b], Laya Mahmoudi [c,d], Jafar Abdollahi [e]

[a] Department of Computer Engineering, University of Zanjan, Zanjan, Iran
[b] Department of Industrial and Systems Engineering, Tarbiat Modares University, Tehran, Iran
[c] Faculty of Economic and Administrative Sciences, Ferdowsi University of Mashhad, Mashhad, Iran
[d] Pardis Cancer Research Center, Pardis Cancer Institute, Shiraz, Iran
[e] Department of Computer Engineering, Faculty of Engineering, Central Tehran Branch, Islamic Azad University, Tehran, Iran

## ARTICLE INFO

*Keywords:*
Multi-modal network
Wound image classification
Wound location
Swin Transformer
Transformer

## ABSTRACT

Effective diagnosis of acute and difficult-to-heal wounds is critical for wound care physicians to provide effective patient care. Poor clinical outcomes are often associated with infection, peripheral vascular disease, and increased wound depth, which collectively exacerbate these comorbidities. This study proposes a multimodal model combining two advanced architectures, a Swin transformer and a Transformer, for wound classification. The multimodal network was built by combining features extracted by the Swin transformer and location features to classify diabetic, pressure, surgical, and venous wound types. The Swin transformer was used to extract image latent features, and the transformer was used to extract location latent features using a decimal encoding map. These features were combined in a fusion layer to adopt the final classification. Swin Transformer and Transformer focuses on wound classification and leverages the strengths of Transformers for a more robust and accurate integration of visual and spatial information. The proposed method was comprehensively compared with deep neural networks (DNNs) for classification on the AZH dataset. Experimental results show significant classification accuracy across wound classes (including only diabetic, pressure, surgical, and venous) in different experiments ranging from 0.7778 to 1.0. The proposed model in four wound class classifications (D vs. P vs. S vs. V) on the AZH dataset with a simplified body map achieved values of 0.8209, 0.8220, and 0.8220 for precision, recall, and F1-score, respectively. The results presented in this study demonstrate the exceptional accuracy of the proposed method in accurately classifying the most common wound types using images of wounds and their respective locations.

## 1. Introduction

Providing internal organs with vital protection from external factors, skin plays an important role in maintaining human health and overall well-being. Skin is highly susceptible to numerous factors that can lead to tissue damage and the onset of trauma (Li et al., 2023). The resulting wounds from the damage can be classified into two categories based on their origin: acute or chronic. Specifically, the acute group consists of injuries resulting from external factors, such as bites, burns, and minor cuts, while chronic wounds stem from internal conditions, such as venous, arterial insufficiency, high blood pressure, and diabetes. Chronic wounds take a lot of time to heal in contrast to acute wounds which heal in a balanced and short period of time (Irfan-Maqsood, 2018). Accordingly, chronic wounds pose a serious health risk due to their severity and burden to the healthcare systems across the world.

The prevalence of chronic wounds in developed countries has been estimated to affect 1 to 2 percent of the population at least once during their lifetime. In a report published by Mission Regional Medical Center in 2020, it was estimated that around 6.7 million people in the United States were suffering from chronic wounds (Maeso et al., 2024), leading to between USD 28–90 billion in medical costs (Saeed & Martins-Green, 2024). Given the significance of this issue, wound diagnosis for effective treatment and management has become a pressing global concern. In traditional methods, wound assessment relies primarily on specialists and is performed manually, often resulting in time-consuming and variable outcomes. To address these issues, the demand for effective diagnostics and management has increased, leading to the advent of artificial intelligence (AI) and significant changes in healthcare. Recent advancements in artificial intelligence (AI) and deep learning have

* Corresponding author.
*E-mail address:* Raminmousa@znu.ac.ir (R. Mousa).

https://doi.org/10.1016/j.eswa.2025.127077
Received 25 September 2024; Received in revised form 21 February 2025; Accepted 25 February 2025
Available online 20 March 2025

shown great promise in the field of medical image analysis. One of the most widely used deep learning models in the tasks of computer vision is the convolution neural network (CNN).

This architecture has shown remarkable results in various computer vision tasks, including object detection or image segmentation (Saha, 2018). However, by focusing primarily on local features in data, CNNs face limitations in capturing long-range dependencies and global context, resulting in reduced performance for tasks that require holistic image interpretation. Accordingly, these challenges have prompted the search for more advanced models such as Transformers, which have demonstrated dominant success in the field of natural language processing (NLP). Inspired by the success of the self-attention-based transformer model in NLP, Dosovitskiy (2020) introduced vision transformer (VIT) architecture for image classification applications (Gheflati & Rivaz, 2022). In a comparison study conducted by Maurício, Domingues, and Bernardino (2023) between CNN and VIT, the authors reported that the ViT architecture demonstrated greater robustness and better performance compared to CNN networks across 17 reviewed papers. Despite the advantages, VIT also faces challenges including quadratic computational complexity and issues with scale variation when adapting Transformer architecture from NLP to computer vision. To address these challenges, Liu et al. (2021) proposed the Swin Transformer, a novel architecture based on shifted window self-attention. This architecture generates multi-scale feature representations and processes images with linear computational complexity, enabling it to handle high dimensional data more efficiently.

Although this technique has shown remarkable results in segmentation, the inherent limitations of CNNs, particularly their difficulty in capturing long-range dependencies and contextual information, have driven the search for more advanced models (Khan et al., 2023). As a result, the Swin Transformer, a novel architecture based on transformer models, has gained attention for its superior performance in image segmentation tasks. Unlike traditional convolutional neural networks (CNNs), the Swin Transformer can capture long-range dependencies in images, making it particularly well-suited for complex medical images. Despite the advancements, the need for more reliable and accurate wound classification results remains a pressing concern.

Deep learning techniques have significantly advanced medical image analysis, particularly wound classification. Convolutional neural networks (CNNs) are widely used due to their ability to automatically extract and learn complex features from wound images and have shown strong performance in various diagnostic tasks. Transfer learning, using pre-trained models such as VGG16, ResNet, and EfficientNet, trained on medical datasets, has been particularly effective in increasing classification accuracy with limited data. However, these models focus primarily on visual features and often ignore spatial information, which is crucial for accurate classification. This visual-centric approach can lead to misclassification due to omitting key location-based contextual cues, such as similar wounds. They can have distinct classifications based on their anatomical location. The inadequate integration of multimodal data, a combination of visual and spatial information, represents a significant gap. To address these limitations, this study proposes a novel multi-faceted approach that integrates image and spatial data to increase the accuracy of wound classification. Our methodology combines Swin Transformer and Transformer, which are specifically designed to process spatial data encoded as binary sequences.

Despite advancements, the need for more reliable and accurate image classification specifically for different wound types remains a pressing concern. Therefore, this study proposes a hybrid approach which combines a Swin Transformer with a transformer. Using a transformer for location data and a Swin Transformer for image data, this approach leverages both architectures' strengths to capture spatial dependencies and complex spatial patterns. The main contributions of this paper can be summarized as follows:

1. To develop and evaluate a multimodal approach for wound classification using a combination of image data with location information.
2. To assess the effectiveness of various machine learning models for wound classification tasks, including convolution neural networks (CNNs) and transformer-based architectures.
3. To investigate the effectiveness of data augmentation on model performance and robustness.

As for the rest of the study, it proceeds as follows. In Section 2, the related works are presented. It is followed by a discussion of the methodology. In Section 3, the materials employed are discussed. In Section 5, the findings of this study are described (see Figs. 1 to 5).

## 2. Related work

In an effort to minimize issues, interest in automated wound assessment procedures has grown over time. The use of swift and accurate systems, processing power, and technological advancements have increased. Researchers continue investigating conventional machine learning algorithms due to their more accessible structure and greater interpretability. An overview of the literature reviewed in this study is presented in Figure Fig. 1. For example, T. Chitra et al. conducted a study utilizing the Random Forest (RF) method to segment and classify wound images and tissues. Their findings demonstrated that the RF method is not only easier to use but also more accurate than manual inspections. The approach is appropriate for physicians at lower levels of telemedicine since it is straightforward, economical, and time-efficient (Chitra, Sundar, & Gopalakrishnan, 2022). Using artificial intelligence, Catarina Pereira et al. created a wound image analysis system that forecasts surgery site infections. The system uses a machine learning classification model to forecast changes and a deep learning segmentation model (MobileNet-Unet) to recognize wound kinds (leg, chest, and drain). While machine learning models use color and texture information to classify wound images, deep learning models segment images and assign wound types. With a mean intersection over the union of 0.899 and a mean average accuracy of 0.901, the segmentation model performed well. With 0.876 recall and 0.526 accuracy, the leg wound classifier produced the best results (Pereira et al., 2023). This study by Syifa'ah Setya Mawarni et al. utilizes GLCM to extract features from wound image test data. The results show that 0.0082 is the contrast value, 0.9769 is the correlation value, 0.6391 is the energy value, and 0.9959 is the homogeneity value. The results of applying the SVM method showed 0.9639 accuracy, 0.9306 precision, 0.9285 recall, and a 0.9258 F1-score. The SVM approach has a 0.9285 classification accuracy for external wound pictures (Murinto, Sunardi et al., 2023). To help with treatment planning, Huang-Nan Huang et al. investigated picture identification of diabetic foot sores. To identify, detect, and size wounds, they employed object recognition technologies, deep neural networks, convolutional neural networks, and the PEDIS score, a qualitative evaluation technique. Medical professionals annotated picture characteristics and trained machine learning modules using the Object Detection Fast R-CNN technique. Researchers discovered that the assessment's accuracy could potentially reach 0.90 (Huang et al., 2022). Takuro Nagataa et al. developed an algorithm that uses machine learning to classify skin tears in digital photos. They used random forest and support vector machine techniques to classify small regions of skin tear images. Thirteen photos were used to assess the effectiveness of these algorithms. The outcomes demonstrated that the algorithms could categorize different skin tears, which might help nurses manage skin tears even without wound care experience. Regarding wound segment and STAR category classification, the support vector machine performed with 0.74 and 0.69 accuracy, respectively (Nagata et al., 2021). According to Zinnel and Bentil (2023), the convolution neural network (CNN) is the most promising machine learning technique for identifying, categorizing, and predicting TBI severity and outcomes in clinical settings.

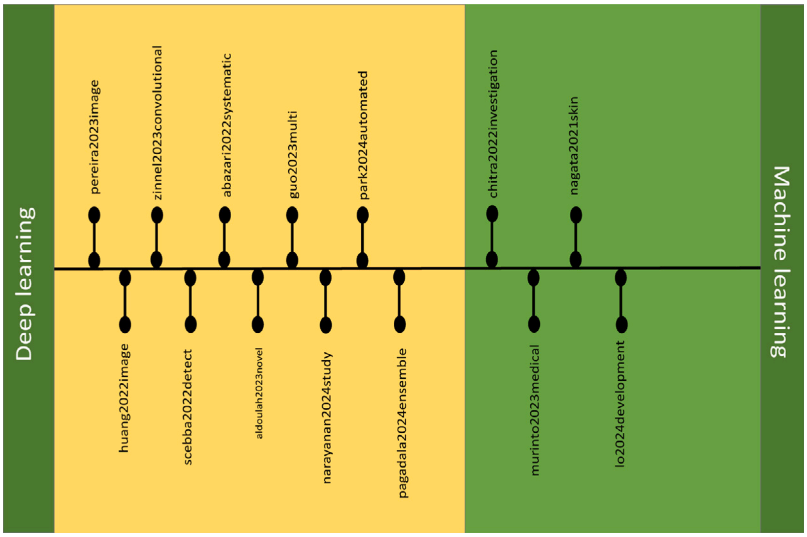


**Fig. 1.** Classification of the reviewed literature in two subsections of machine learning and deep learning.

Scebba et al. (2022) developed Detect-and-Segment (DS), a deep learning method for image-based wound diagnosis. They used deep neural networks to locate the wound, separate it from the background, and create a segmentation map. They evaluated the DS using four separate data sets, including data on diabetic foot ulcers. With the ulcer data set, Matthews' correlation value increased from 0.29 to 0.85. The DS can train segmentation models with up to 0.90 less training data without compromising performance. This strategy, which automates wound analysis, aims to reduce the amount of work required to maintain chronic wounds. A study by Huang et al. (2023) introduced a CNN-based model in wound assessment aimed at assisting non-specialist medical personnel. Using a multi-task deep learning framework, their model could simultaneously classify five key wound types: deep, infected, arterial, venous, and pressure wounds. The findings showed that the proposed model outperformed or matched the performance of medical personnel across five wound classification tasks.

Abazari, Ghaffari, Rashidzadeh, Badeleh, and Maleki (2022) systematically reviewed studies focusing on burn wound classification, identification, and healing processes. This review conducted a detailed analysis of 16 articles on wound classification, 52 on wound healing, and 18 on identifying and classifying new wounds. As a result, this review highlights the need for a more comprehensive classification system for burn wounds, considering more effective and critical factors that fall into two main categories: local factors and systemic factors.

Lo et al. (2024) proposed an explainable AI model in their study to analyze 2957 wound images taken from the Singapore Vascular Wound Registry. The AI system demonstrated proficiency across multiple tasks, achieving 0.959 accuracy in wound classification, 0.850 in in-depth assessment, 0.871 in width and length determination, and 0.878 in wound segmentation. Notably, explainable AI techniques were employed in this research to achieve high accuracy and enhance the transparency of the developed model as a potential model, making it a potential tool for improving wound assessment among Asian populations.

By addressing a substantial gap in the existing literature, which primarily focuses on binary outcomes, Aldoulah, Malik, and Molyet (2023) introduced a novel deep learning framework called Swish-ELU EfficientNet-B4 (SEEN-B4), capable of classifying chronic wounds into multiple categories. Their proposed model was implemented on publicly accessible datasets, including Medetec and AZH datasets as well as their extended version,to tackle issues associated with class imbalance. Compared to existing state-of-the-art methods, the SEEN-B4 model achieved superior accuracy rates of 0.8732, 0.8817, 0.88, and 0.8934 on the AZH, extended AZH, Medetec, and extended Medetec datasets, respectively. These high accuracy rates indicate the model's robustness across different datasets in addition to its effectiveness in classifying chronic wounds. To enhance the accuracy of multi-class wound classification, Guo et al. (2023) addressed the challenges posed by the complexity and variety of wound images, which render thewidely used deep learning model, Convolutional Neural Network (CNN), inefficient during feature extraction. Accordingly, the authors presented a novel High and Low-Frequency Guidance Network (HLG-Net), consisting of two branches, namely the High-Frequency Network (HF-Net) and the Low-Frequency Network (LF-Net), to extract detailed textures and global information from wound images, respectively. By combining the extracted features from both branches, the architecture demonstrated impressive accuracies of 0.9800 for two-class, 0.9211 for three-class, and 0.8261 for four-class classifications.

Another study conducted by Narayanan and Ghanta (2024) addressed the critical challenge of training data scarcity, which arises primarily due to privacy and legal concerns. To overcome this challenge, the authors employed two main data augmentation approaches: geometric transformations and Generative Adversarial Networks (GANs). Using state-of-the-art computer vision models (MobileNet V2, ResNet50, and VGG16) as a baseline, they demonstrated that geometric data augmentation, including rotating and brightening images, improved classification by up to 0.11 in F1-scores across key wound categories. Despite the success of their experiments with DE-GANs, they concentrated primarily on diabetic ulcers due to computational constraints and mode collapse challenges. Their findings revealed that generating synthetic wound images with rich variations using DE-GANs did not consistently translate into improved classification accuracy.

In addition to the need to accurately classify wound types, Park and Sung (2024) highlighted the importance of rapid wound assessment in emergency situations and quick decision-making about wound treatment. To address this, the researchers employed two state-of-the-art image classification models: ResNeXt and Vision Transformer (ViT). By training on a combined dataset of about 1000 images from the Medetec and AZH datasets, the proposed ViT-based modeloutperformed the previous studies in this field, achieving remarkable performance metrics with limited data: 0.9278 accuracy, 0.9489 precision, 0.9187 recall, and 0.9244 F1-score. Another Vision Transformer (ViT)-based study, conducted by Pagadala, Silas, and Joy (2024), focused on the accurate and efficient classification of Diabetic Foot Ulcers (DFU), specifically distinguishing between healthy skin and ulcerous tissue. To achieve this, the authors developed an ensemble model combining two different architectural approaches: ResNet50 with Vision Transformers and MobileNet with Vision Transformers. When tested on a dataset

**Table 1**
A summary of recent studies on wound pictures.

| Ref | Data type | Target output | Num. of samples | Method | Performance measure |
|---|---|---|---|---|---|
| Chitra et al. (2022) | 3-D representation | Classify tissues and segment wound images | | Random Forest (RF) | 0.938 |
| Pereira et al. (2023) | RGB images | Predict wound changes, identify the wound area | 1443 | MobileNet-Unet and KNN and RF | 0.901 |
| Murinto et al. (2023) | Wound Dataset | Severity of a diabetic foot ulcer | 280 | SVM algorithm | and 0.9285 |
| Huang et al. (2022) | DFUC 2020 Dataset | Image segmentation | 3600 | Fast R-CNN | 0.90, |
| Nagata et al. (2021) | JPEG format | Skin Tear classification | 31 images | SVM classification | 0.74 and 0.69 |
| Zinnel and Bentil (2023) | brain imaging or EEG data | Diagnosis and management of skin tears | | convolutional neural network | – |
| Scebba et al. (2022) | Medetec Medical Images | Wound segmentation | 1096 | Deep Learning (DL) | 0.90 |
| Huang et al. (2023) | Color wound images | Binary classification | 2149 wound images | DL-based CNN model | Accuracy,<br>Sensitivity,<br>Specificity,<br>AUC,<br>Kappa |
| Lo et al. (2024) | Asian Vascular wound images | Wound classification (4 types)<br>Wound measurements (width, length, depth)<br>Wound segmentation (18 features) | 2957 wound images | DenseNet, MobileNet, and ResNet for classification<br>DeepLab, FPN, U-Net for segmentation | Accuracy,<br>F1-score,<br>AUROC,<br>Confidence,<br>Explain ability scores |
| Chitra et al. (2022) | Chronic wound tissue images | Classification of wound tissues into three categories | Not specified | Random Forest (RF) algorithm | Accuracy |

of more than 1000 images, ResNet50-ViT achieved 0.9375 validation accuracy, while MobileNet-ViT reached 0.9688 (See Table 1).

## 3. Methodology

While transformer architecture is highly demanded for its exceptional performance in natural language processing (NLP) tasks, its application has expanded significantly as a general-purpose backbone in computer vision (CV). However, the transfer from NLP to CV introduces two main challenges, reflecting the difference between language and vision data: (1) object detection, and (2) semantic segmentation. Compared to NLP, which uses a relatively fixed scale for tokens, visual elements in images can vary greatly in scale. Similarly, because images generally have higher resolutions compared to text, tasks involving dense predictions, such as semantic segmentation, are complicated by the quadratic computational complexity of traditional transformers. Consequently, the first transformer model to utilize purely self-attention, known as the Vision Transformer (ViT), was introduced for image recognition by Dosovitskiy (2020). Although ViT has demonstrated outstanding performance in computer vision task (Ayas & Tunc-Gormus, 2022), it struggles with variations in object size within images and the high resolution of images, which exceeds the capacity of the model (Chen, 2022). To overcome these challenges, a ViT-based model called the Swin Transformer was developed, which constructs hierarchical feature maps to manage the different scales of visual elements effectively. This model starts with small patches for the first transformer layer. Then it merges them into bigger ones in the deeper transformer layers, leading to the building of more abstract and hierarchical representations of the image. To address the high complexity of applying transformers to high-resolution images, the Swin Transformer leverages computing self-attention locally within non-overlapping windows of the image.

### 3.1. Method

Similar to the vision transformer (ViT), the initial processing in the Swin transformer starts with splitting the input RGB image into non-overlapping patches. The features of each patch are represented by the combination of raw pixel RGB values, followed by their projection into an arbitrary dimension through a linear embedding layer. A modified self-attention mechanism with a window-based approach is applied to achieve linear computational complexity. More specifically, in contrast to the traditional transformers which are impractical in processing high-resolution images, Swin Transformer addresses this issue by dividing the inputs into non-overlapping windows and then computing the self-attention within each window. The computations for both Traditional and Swin transformers are as follows:

$$\Omega(MSA) = 4hwC^2 + 2(hw)^2C \tag{1}$$

$$\Omega(W-MSA) = 4hwC^2 + 2M^2(hw)C \tag{2}$$

While global self-attention computes the relationships between all pairs of patches, the Swin Transformer reduces computational complexity by considering self-attention independently within each window. In this method, the computational complexity is $2M^2hwC$ and scales linearly with the number of windows, rather than quadratically as represented by $2(hw)^2C$ in Global Self-Attention. Despite reducing the computation, the window-based self-attention module has a limitation in cross-window connections, addressed by alternating between two window partitioning configurations—regular and shifted—in successive Swin Transformer blocks.

In the first step, the module applies a regular window-based self-attention (W-MSA) from the top-left pixel of the image and divides an $8*8$ feature map into two $2*2$ windows, each $4*4$.

$$\hat{z}^l = W-MSA(LN(z^{l-1})) + z^{l-1} \tag{3}$$

$$z^l = MLP(LN(\hat{z}^l)) + \hat{z}^l \tag{4}$$

Mathematically, the formulas refer to a two-process computation applied in the Swin transformer using window-based multi-head self-attention (W-MSA) and a multi-layer perceptron (MLP). In the first process, the normalization layer (LN) is performed on the input features, obtained from the previous layer ($Z^{l-1}$). Using normalized features, W-MSA calculates self-attention for each non-overlapping window. Compared to global self-attention, localized attention shows a significant reduction in computational complexity due to considering the relationships within each window. To ensure that the original input features are preserved, the residual connection $+z^{l-1}$ is added to the output of W-MSA $LN(Z^{l-1})$. For the next step, the normalization layer is applied once more on the output features from the previous step, denoted as $z^l$. This step is followed by feeding the processed output into a multi-layer perceptron (MLP). The MLP applies a sequence of linear transformations and non-linear activations, improving the feature representation. To ensure the continuity and stability of the process in the learning, the original W-MSA features through a residual connection, $+z^l$ are combined with the MLP output.

$$\hat{z}^{l+1} = SW-MSA(LN(z^l)) + z^l \tag{5}$$

$$z^{l+1} = MLP(LN(\hat{z}^{l+1})) + \hat{z}^{l+1} \tag{6}$$

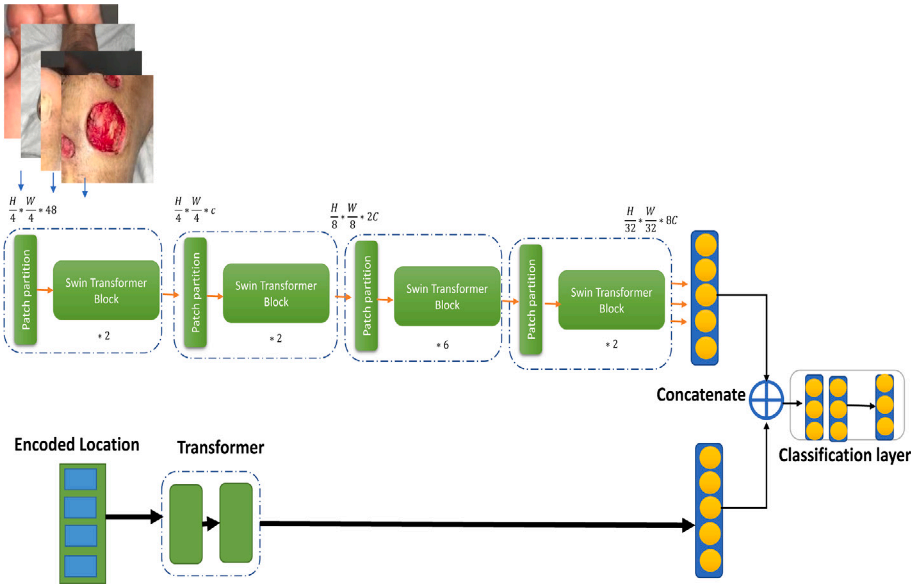


**Fig. 2.** An overview of proposed multi-modal wound classifier.

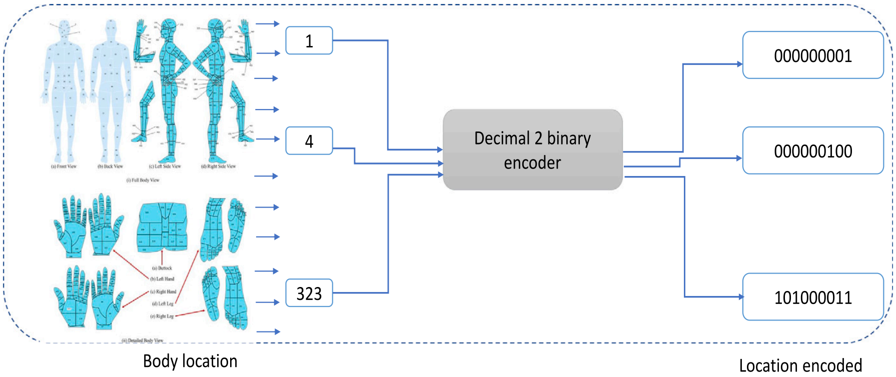


**Fig. 3.** The process of converting numbers related to the location of wounds from decimal to binary.

To address the window-based self-attention mechanism's inability to perform cross-window connections, shifted window-based multi-head self-attention (SW-MSA) is applied to the previous layer's configuration. To this end, the normalization layer is initially applied to the previous MLP output features and then passed through the Shifted Window-based Multi-head Self-Attention mechanism. The residual connection $+z^l$ is added to the SW-MSA output to help retain the original input features. Again, the output feature of $SW-MSA(LN(z^l))+z^l$, denoted as $z^l+1$, is normalized by the normalization layer $(LN(\hat{z}^{l+1}))$. Following the normalization process, another MLP is applied to the output features to enhance their representation. The MLP output combined with the residual connections, $\hat{z}^{l+1}$, involving the original features from the SW-MSA step. Despite the advantages, an issue that arises when implementing Shifted Window Partitioning is the increase in the number of windows, which can lead to inefficiencies. More specifically, the initial configuration has a total of $(\lceil \frac{h}{M} \rceil * \lceil \frac{w}{M} \rceil)$ windows, while the shifted configuration increases this number to$(\lceil \frac{h}{M} \rceil + 1) * (\lceil \frac{w}{M} \rceil + 1)$, causing some windows to become smaller than $M * M$ in size. As an instant solution, padding is applied to the smaller windows to reach the full size of $M * M$. This approach increases the computation, even though it masks out the padded values during attention computation. To address this issue, Liu et al. (2021) propose a more efficient batch computation method, which entails shifting windows cyclically in the direction of the top-left. In this method, the edge going beyond the shifted windows wraps around the opposite edge. As a result, a batched window is created where sub-windows are not positioned next to each other in the original feature map. This lack of adjacency requires the implementation of a masking mechanism to maintain efficiency in self-attention computations. An overview of Swin Transformer application on the input dataset is given in Fig. 2. The output of applying Swin Transformer on the input images is called $Image_{latent}$. This vector contains the low-level features extracted from the image.

### 3.2. Transformer

The proposed transformer-based approach for extracting features from the disease location is shown in Fig. 4. Before examining the details of the transformer, we will first pre-process the data related to the location of the wounds. These features, originally numerical values ranging from 1 to 323, are converted into binary code. The decimal to binary conversion process is shown in Fig. 3. In fact, a 9-digit binary number was considered for each location of the disease. This 9-digit characteristic was considered as the input of the transformer in the form of the sequence $Loc = [X_0, X_1, \dots, X_8]$.

Let $S = X_i, y_{i}{}_{i=1}^{r}$ denote a set of problem inputs(binary numbers). Where $X_i$ represents the input signal, and $y_i$ represents the associated class. $y \in 1, 2, 3, \dots, m$ where $m$ is the number of classes defined in the data set. The Transformer model aims to learn the mapping from the sequence $X$ to the corresponding semantic label. The proposed transformer is completely based on the vanilla Transformer (Cheng

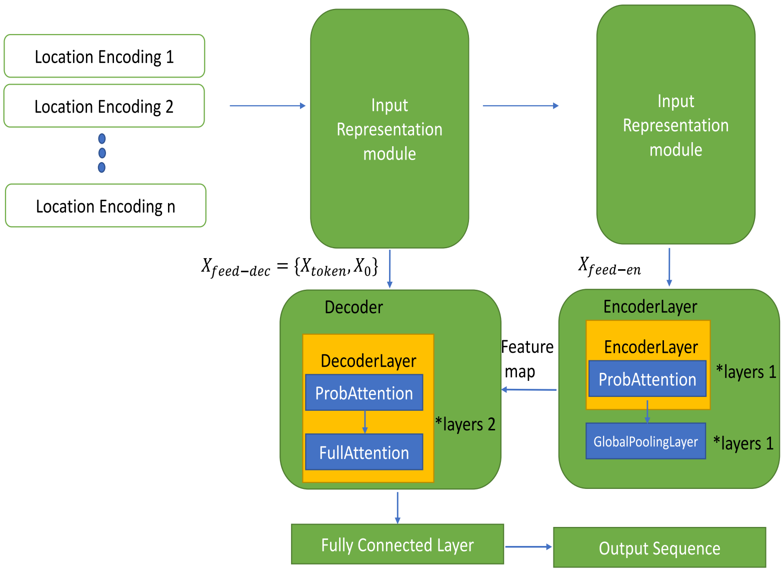


**Fig. 4.** An overview of the transformer model for binary decoded data.

et al., 2024), an architecture that has attracted much attention in recent years by showing improved performance in machine translation and other NLP tasks. The transformer follows an encoder–decoder architecture that can process sequential data in parallel without relying on redundant networks. The success of transformer models has primarily benefited from the self-attention mechanism, which has been proposed to capture long-range relationships between sequence elements. Vanilla Transformer is proposed as an attempt to extend the use of standard transformers to time series classification. Unlike conventional CNN and RNN architectures that typically use filters with a local receptive field or input sequence order, the attention mechanism employed by Vanilla Transformer allows it to focus different input regions. The complete architecture of the whole model is shown in Fig. 4. It consists of an embedding layer, an encoder and a final classifier. The first step transforms an input $X$ from the training set into rolling windows. The transformer considers each window as one view. Therefore, the input $X$ is represented in the form $T * H * FT * H * F$, where $T$ indicates the number of previous time steps, $H$ indicates the row of data (its value is considered 1), and $F$ indicates the number of features in the data. Below are the components of the proposed transformer.

### 3.2.1. Linear embedding layer

Before feeding the sequence of rolling windows to the encoder, the input is linearly predicted in the dimension vector of the model $d$ using the learned embedding matrix $E$. The embedded representations are then concatenated together with a learnable classification token $v_c$ required to perform the classification task. The transformer views embedded inputs as a collection of patches in no particular order. To maintain the spatial arrangement of the patches as in the original sequence, the positional information $E_p$ is encoded and added to the rolling window displays. The embedded sequence resulting from rolling windows is given by:

$$z_{=}[v_{class}; x_1 E; x_2 E; ...; x_n E] + E_p os \tag{7}$$

Where $E \in \mathbb{R}^{(p^2 c)*d}$ and $E_{pos} \in \mathbb{R}^{(n+1)*d}$.

It is claimed in Dosovitskiy (2020) and Bazi, Bashmal, Rahhal, Dayil, and Ajlan (2021) that $1 - D$ and $2 - D$ positional encoding produce almost identical results. Therefore, a simple $1 - D$ positional encoding is used in the proposed model to maintain the positional information.

### 3.2.2. Transformer encoder

The embedded sequence $z_0$ is sent to the Transformer encoder. The encoder can consist of L identical layers (in the proposed model, it consists of one layer). Each layer has two main components:

1. Multihead self-attention block (MSA):

$$z'_\iota = MSA(LN(z_{\iota-1})) + z_{\iota-1}, \iota = 1, \dots, l \tag{8}$$

2. Fully connected feed-forward dense block (MLP):

$$z_\iota = MLP(LN(z'_\iota)) + z'_\iota, \iota = 1, \dots, l \tag{9}$$

LN stands for Normalization layer. LN is a function that maps $\mathbb{R}^D \to \mathbb{R}^D$ using two parameters: gains $\alpha$ and biases $\beta$. This relationship is as follows:

$$LN(z; \alpha, \beta) = \frac{(z - \mu)}{\sigma} \odot \alpha + \beta \tag{10}$$

$$\mu = \frac{1}{D} \Sigma_{i=1}^{D} z_i$$

$$\sigma = \sqrt{\frac{1}{D} \Sigma_{i=1}^{D} (z_i - \mu)^2}$$

Where, $z_i$ is the $i$th element of the vector $z$.

In the last encoder layer, we take the first element in the sequence $z_0^L$ and send it to an external head classifier to predict the class label.

$$y = LN(z_0^L) \tag{11}$$

The MSA block in the encoder is the central component of the transformer. It determines the relative importance of embedding a sequence with respect to other embeddings in the sequence. This block has four layers:

1. linear layer
2. Self-attention
3. Concatenation layer
4. Final linear layer

The self-attention score was found to have a high "scatter", meaning that some dot-product pairs could contribute to the attention mechanism, and others could be ignored (Zhou et al., 2021). The $i - h$ query attention can be defined as a kernel in the following form:

$$Attention(q_i, K, V) = \sum \frac{k(q_i, k_j)}{\sum k(q_i, k_l)} v_j \tag{12}$$

In this regard, the attention of query $i$ to all keys is defined as probability $p(k_j|q_i)$, and the output is its combination with $v$ values. We want to identify the most important queries that can be achieved by measuring the similarity between p and q using the Kullback–Leibler divergence. The dispersion measure of query $i$ can be defined as follows:

$$M(q_i, K) = ln \sum_{j=1}^{l_k} e^{\frac{q_i K^T}{\sqrt{2}}} - \frac{1}{L_K} \sum_{j=1}^{l_k} e^{\frac{q_i K^T}{\sqrt{2}}} \tag{13}$$

In the following the probsparse self-attention can be defined as:

$$Attention(Q, K, V) = softmax(\frac{\hat{Q} K^T}{\sqrt{d}})V \tag{14}$$

Given an input sequence of tokens represented as a matrix $X$ with dimensions (sequence_length, embedding_dimension), the attention mechanism calculates a set of attention scores $A$ as follows:

For each position $i$ in the sequence:

1. Generate three new matrices:
   - Query matrix $Q_i$ by multiplying $X_i$ (the input at position $i$) with a learnable weight matrix $W_Q$.
   - Key matrix $K_i$ by multiplying $X$ with another learnable weight matrix $W_K$.
   - Value matrix $V_i$ by multiplying $X$ with a third learnable weight matrix $W_V$.

2. Compute the attention scores between the query $Q_i$ and all key positions in the sequence using the dot product:

$$A_i = \text{softmax}\left(\frac{Q_i K^T}{\sqrt{d_k}}\right) \tag{15}$$

   - $A_i$ represents the attention scores for position $i$.
   - $Q_i$ is the query matrix for position $i$.
   - $K^T$ is the transpose of the key matrix.
   - $d_k$ is the key vectors' dimension, typically a fraction of the embedding dimension.

3. Use the attention scores to compute a weighted sum of the value matrices:

$$O_i = A_i V \tag{16}$$

   - $O_i$ represents the output (context) vector for position $i$.
   - $A_i$ is the attention scores for position $i$.
   - $V$ is the set of value matrices for all positions.

The attention scores reflect how much each position's information contributes to the representation of the current position. This mechanism allows the model to focus more on relevant parts of the input when making predictions or encoding information.

In summary, the attention mechanism calculates attention scores by comparing queries with keys to measure the importance of different positions in the input sequence. This process captures relationships and dependencies, enabling the model to understand context and relationships within the data.

The attention mechanism can be mathematically represented as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{17}$$

The Attention Calculation Formula is as follows:

$Q$ : Query matrix

$K$ : Key matrix

$V$ : Value matrix

$d_k$ : Dimension of keys

$n$ : Number of elements in the sequence

The attention $score$ between a query element $q_i$ and a key element $k_j$ is calculated as:

$$score(q_i, k_j) = q_i \cdot k_j^T \tag{18}$$

To improve the weighting, the attention scores are often scaled by the square root of the dimension of keys $d_k$:

$$scaled\_score(q_i, k_j) = \frac{score(q_i, k_j)}{\sqrt{d_k}} \tag{19}$$

The scaled scores are then passed through a Softmax function to get the attention weights $w_{ij}$:

$$w_{ij} = \frac{e^{scaled\_score(q_i,k_j)}}{\sum_{j=1}^{n} e^{scaled\_score(q_i,k_j)}} \tag{20}$$

These attention weights represent the importance of each key element $k_j$ with respect to the query element $q_i$. They are then used to compute a weighted sum of the value elements $v_j$ to obtain the final output:

$$attention(q_i, K, V) = \sum_{j=1}^{n} w_{ij} \cdot v_j \tag{21}$$

### 3.3. Encoder and decoder

The proposed model uses the encoder module to capture the inputs long-term dependence. The $t^{\text{th}}$ input $X_t$ is mapped to a matrix $X^t_{feed_en} \in R^{Lx} * d_{model}$. The encoder consists of several attention layers and Global Pooling layers. We use a distillation operation to select the top $V$ value compounds with dominant properties. The procedure between the two layers is defined by the following relationship:

$$X^t_{j+1} = ELU(GlobalPooling([X^t_j]_{AB})) \tag{22}$$

in this equation $[.]_{AB}$ is the multi-head ProbSparse self-attention. The feature map generated by the encoder is fed to the decoder. The receiver has two identical multihead attention layers. To reduce the speed loss in long prediction, in the proposed model, the following vector is entered into the encoder:

$$X^t_{feed_de} = Concat(X^t_{token}, X^t_0) \tag{23}$$

where $X^t_{token}$ is embedded and time stamp and $X_0$ is a placeholder for target sequence. The output of applying transformer on the input encoded location is called $Location_{latent}$. This vector contains the low-level features extracted from the encoded location.

At the end, the combination of $Image_{latent}$ and $Location_{latent}$ are combined in a concatenate layer. The following is how to combine these two vectors:

$$Finall_{vector} = Image_{latent} \bigoplus Location_{latent} \tag{24}$$

Finally, the $Finall_{vector}$ is used for classification. After passing through several layers, this combination is mapped to an output layer, which has $n$ neurons ($n$ number of data classes) and the Softmax function is used to calculate the probability of classes.

In this study, the fusion of image and positional features is achieved through a straightforward concatenation of latent vectors extracted from each modality. Specifically, the $Image_{latent}$ vector is derived from the Swin Transformer, which processes wound images by dividing them into non-overlapping patches, projecting them into a higher-dimensional space, and applying a window-based self-attention mechanism. This results in a vector representation of the image's low-level features. Concurrently, the $Location_{latent}$ vector is generated from encoded wound location data, transformed into a 9-digit binary sequence and processed by the Transformer model, which captures spatial patterns and relationships. These two latent vectors, $Image_{latent}$ and $Location_{latent}$, are combined in a concatenation layer to create the $Finall_{vector}$. This fused vector serves as input to the classification layers, which consist of several processing layers culminating in a Softmax activation function to predict the probability of each wound class. While the current fusion method effectively integrates features from both modalities, it does not explicitly address the relative importance of image and location features. Advanced strategies such as weighting schemes or attention mechanisms were not employed, leaving room for further exploration in balancing feature importance to prevent potential information loss. This approach, however, demonstrates a robust framework for combining multimodal data, ensuring consistency and simplicity in feature integration.

## 4. Material

### 4.1. Dataset: AZH Wound and Vascular Center Database

In this study, we use an open-source wound dataset, the AZH dataset, containing 730 images of four distinct types of wounds, namely venous, diabetic, pressure, and surgical wounds. To collect this dataset, a two-year clinical effort was undertaken at the wound care center in Milwaukee, Wisconsin, USA. Afterward, each image was assigned to a wound type by a specialist from the center. As a leading center, the AZH Wound and Vascular Center provides specialized services to treat and manage chronic and complex wounds. The wound images were taken using high-quality imaging devices, including an iPad Pro (with software version 13.4.1) and a Canon SX 620 HS digital camera. These images were stored in JPEG format, with sizes ranging from 320 to 700 pixels in width and 240 to 525 pixels in height. This dataset includes only four types of wounds and surrounding skin, with each image primarily representing a single patient, excluding non-essential information such as personal particulars. Among the cases, some image data were captured from different body sites of the same patient or at various stages of wound healing, but they were considered separate entries in the dataset due to their unique features. Importantly, all images in the AZH dataset were anonymized by the authors, ensuring that no patient-identifying information was present. Notably, no additional samples were added to the public dataset we were working on due to the challenges of capturing more images under controlled clinical conditions. This dataset includes neither any direct experiments conducted on humans nor the use of human tissue samples. It can be accessed online at the following GitHub repository: https://github.com/uwm-bigdata/Multi-modal-wound-classification-using-images-and-locations.

### 4.2. Deep learning library

In this research, Keras was used to implement neural networks. Keras[1] is an open-source machine learning library that has recently become very popular for implementing deep learning models. Keras has a simple and intuitive interface for developing deep neural network models, enabling implementation by calling layers and functions. Additionally, this library supports multiple backends, such as Tensorflow[2] and Theano,[3] reducing the complexity threshold for developing neural network models. In general, this framework is known as a high-level user interface. Table 2 shows the hardware and software specifications of this research.

**Table 2**
The hardware and software specifications of this research.

| Software | | |
|---|---|---|
| Name | version | Description |
| Ubuntu Bionic Beaver (LTS) | 18.04.2 | Operating System |
| Python | 3.6.7 | Used for implementation |
| Keras | 2.2.4 | Used for building models |
| Pandas | 0.23.4 | Used for data analysis |
| Tensorflow | 2.12.0 | Used as backend for Keras |
| CUDA | 9.0.176 | Required for Tensorflow |
| cuDNN | 7.4.1 | Required for Tensorflow |
| Hardware | | |
| Name | Version | |
| CPU | Intel i7-2600 | |
| GPU NVIDIA | GeForce GTX 980 | |
| Memory | Kingston 8 GB DDR3 | |
| GPU Memory | 4 GB, GDDR5 | |

**Table 3**
Hyperparameter setting of the proposed model.

| Hyperparameter | Values |
|---|---|
| Batch size | 64 |
| Learning rate | 0.00005 |
| Dropout | 0.5 |
| Epochs | 200 |
| Optimizer | Adam |
| Loss function | Multi class classification (CrossEntropyLoss)<br>Binary class classification (BCELoss) |

### 4.3. Evaluation metrics

We used the following evaluation metrics to assess the performance of our proposed model: accuracy, precision, recall, F1-Score, and specificity.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{25}$$

$$Precision = \frac{TP}{TP + FP} \tag{26}$$

$$Recall = \frac{TP}{TP + FN} \tag{27}$$

$$F1 - score = 2 * \frac{precision * recall}{precision + recall} \tag{28}$$

$$Specificity = \frac{TN}{TN + FP} \tag{29}$$

## 5. Results

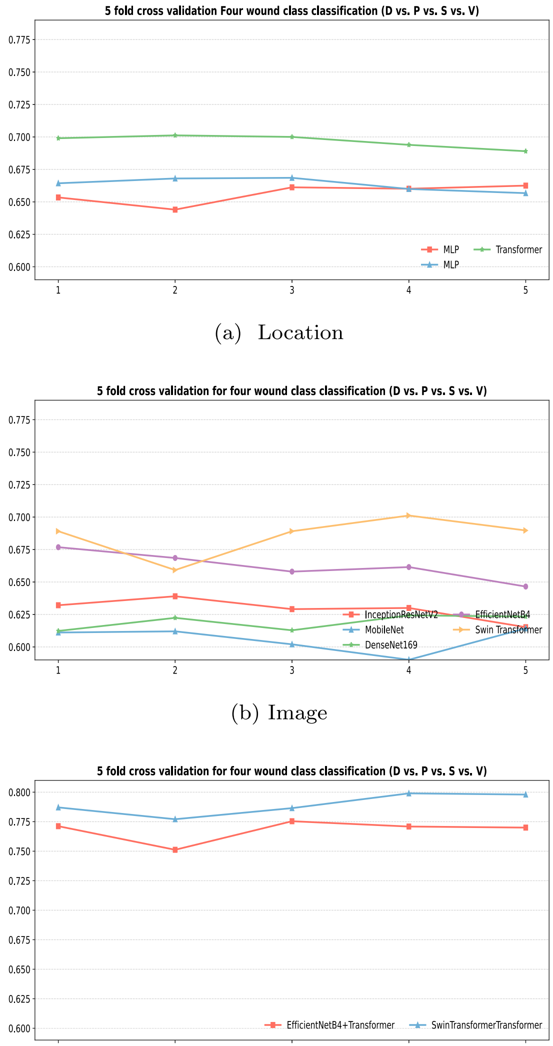
The datasets are presented in two separate parts: train and test. The best results achieved by the tested models were recorded in fold-5. Fig. 5 shows the results obtained by the models in four-class wound classification (D vs. P vs. S vs. V). The best results obtained by the tested models were in fold-5. The configuration of the proposed model and its hyper-parameters are summarized in Table 3, respectively.

Table 4 compares the proposed model and baseline models for the four-class wound classification task (D vs. P vs. S vs. V) on the AZH dataset with the original body map. The models were evaluated on both the original and augmented datasets, focusing on three types of inputs: location, Image, and their combination. The Transformer model consistently outperformed the MLP and LSTM baselines for location data, with its accuracy increasing from 0.6924 to 0.7474 after data augmentation. In addition to the approaches reviewed in Anisuzzaman et al. (2022), four other approaches were

[1] https://keras.io/.

[2] https://www.tensorflow.org/.

[3] https://pypi.org/project/Theano/.



(a) Location

(b) Image

(c) Image and Location

**Fig. 5.** 5-fold cross-validation on four-class wound classification (D vs. P vs. S vs. V).

used to compare the proposed approach on image data, including InceptionResNetV2 (Szegedy, Ioffe, Vanhoucke, & Alemi, 2017), MobileNet (Howard, 2017), DenseNet169 (Huang, Liu, Van Der Maaten, & Weinberger, 2017), and EfficientNetB4 (Pillai, Sharma, & Gupta, 2023). These approaches were trained on the same input dataset and imageNet weights. Image-based models, including VGG19, ResNet, and InceptionV3, demonstrated suboptimal performance in their original configurations. The InceptionResNetV2 model was able to achieve an accuracy of 0.6391 in the original image data and an accuracy of 0.6717 in the augmented mode. Better results were also obtained in the augmented mode for other comparative approaches. The proposed Swin Transformer approach achieved an accuracy of 0.6791 in this data. This model also achieved an accuracy of 0.7223 in the augmented mode, the highest among image-based input methods. In the Image + Location data combination mode, the models examined in Anisuzzaman et al. (2022) achieved a maximum accuracy of 0.7717. Also, the EfficientNetB4+ Transformer model achieved an accuracy of 0.7712 in the original data. The Swin Transforme + Transforme achieved an accuracy of 0.7871 in this data and an accuracy of 0.8189 in the augmented mode. Other evaluation metrics of the approaches are also mentioned in the table. Also, the bar plot for the four-class classification in both original and augmented data is shown in Figs. 6 and 7.

Table 5 compares the performance metrics (accuracy, precision, recall, and F1-score) of various machine learning models using original and augmented data. It evaluates models based on location data as well as a combination of image and location data. Key findings include:

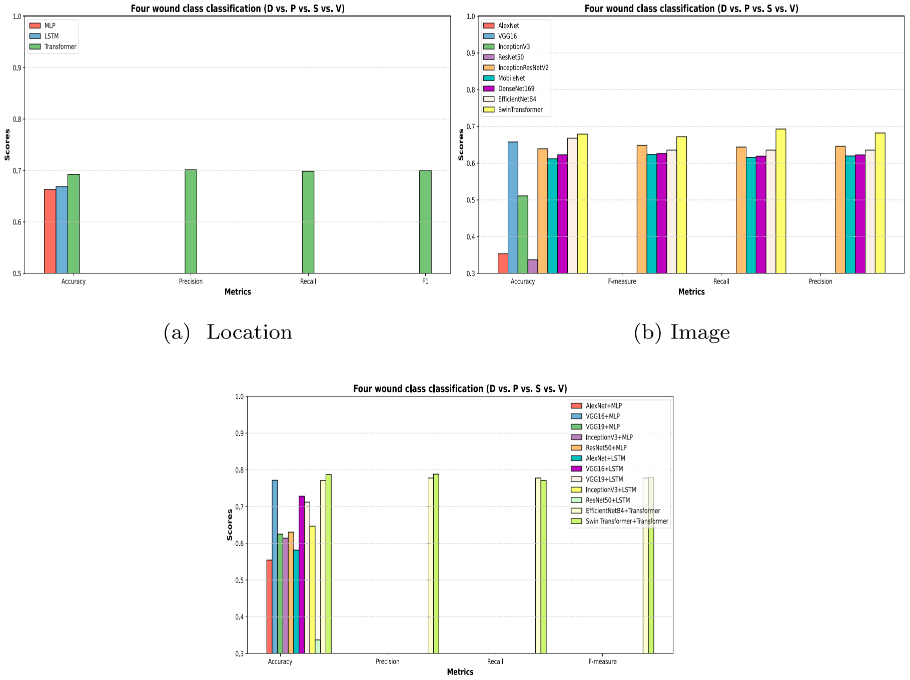
1. Location Data Models: With original data, the Transformer model achieved the highest accuracy (0.7423), then improved to 0.7689 with augmented data.
2. Image + Location Data Models: The VGG19 + LSTM model showed the highest accuracy with original data (0.7935), while the Swin Transformer + Transformer model demonstrated the best performance with augmented data, achieving an accuracy of 0.8312 and an F1-score of 0.8220.
3. When using augmented data, most models improved in accuracy and other metrics, highlighting the effectiveness of data augmentation in enhancing model performance.

Table 6 compares the accuracy of various machine learning models using different types of input data (location, image, and a combination of both). Key findings include:

1. Location Data Models: The Transformer model achieves the highest accuracy (0.6772).
2. Image Data Models: With an accuracy of 0.7800, the Swin Transformer outperforms other models.

**Table 4**
Four wound class classification (D vs. P vs. S vs. V) on AZH dataset with original body map.

| | Models | Accuracy | Precision | Recall | F1-score | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|---|---|---|---|
| | | | Original Data | | | | Augmented data | | |
| Location | MLP (Anisuzzaman et al., 2022) | 0.6630 | – | – | – | 0.7174 | – | – | – |
| | LSTM (Anisuzzaman et al., 2022) | 0.6685 | – | – | – | 0.7228 | – | – | – |
| | Transformer | 0.6924 | 0.7015 | 0.6986 | 0.7000 | 0.7474 | 0.7472 | 0.7448 | 0.7459 |
| Image | AlexNet (Anisuzzaman et al., 2022) | 0.3533 | – | – | – | 0.3750 | – | – | – |
| | VGG16 (Anisuzzaman et al., 2022) | 0.6576 | – | – | – | 0.7173 | – | – | – |
| | VGG19 (Anisuzzaman et al., 2022) | 0.5652 | – | – | – | 0.6304 | – | – | – |
| | InceptionV3 (Anisuzzaman et al., 2022) | 0.5109 | – | – | – | 0.5609 | – | – | – |
| | ResNet50 (Anisuzzaman et al., 2022) | 0.3370 | – | – | – | 0.3370 | – | – | – |
| | InceptionResNetV2 | 0.6391 | 0.6486 | 0.6438 | 0.6461 | 0.6717 | 0.6806 | 0.6736 | 0.6770 |
| | MobileNet | 0.6120 | 0.6237 | 0.6157 | 0.6196 | 0.6500 | 0.6653 | 0.6586 | 0.6619 |
| | DenseNet169 | 0.6228 | 0.6262 | 0.6190 | 0.6225 | 0.6174 | 0.6247 | 0.6038 | 0.6140 |
| | EfficientNetB4 | 0.6680 | 0.6355 | 0.6355 | 0.6355 | 0.7152 | 0.7278 | 0.7224 | 0.7250 |
| | Swin Transformer | 0.6791 | 0.6718 | 0.6929 | 0.6821 | 0.7223 | 0.7766 | 0.7738 | 0.7751 |
| Image + Location | AlexNet + MLP (Anisuzzaman et al., 2022) | 0.5543 | – | – | – | 0.6141 | – | – | – |
| | VGG16 + MLP (Anisuzzaman et al., 2022) | 0.7717 | – | – | – | 0.78 | – | – | – |
| | VGG19 + MLP (Anisuzzaman et al., 2022) | 0.6250 | – | – | – | 0.7228 | – | – | – |
| | InceptionV3 + MLP (Anisuzzaman et al., 2022) | 0.6141 | – | – | – | 0.711 | – | – | – |
| | ResNet50 + MLP (Anisuzzaman et al., 2022) | 0.6304 | – | – | – | 0.6685 | – | – | – |
| | AlexNet + LSTM (Anisuzzaman et al., 2022) | 0.5815 | – | – | – | 0.6685 | – | – | – |
| | VGG16 + LSTM (Anisuzzaman et al., 2022) | 0.7283 | – | – | – | 0.7935 | – | – | – |
| | VGG19 + LSTM (Anisuzzaman et al., 2022) | 0.71200 | – | – | – | 0.7663 | – | – | – |
| | InceptionV3 + LSTM (Anisuzzaman et al., 2022) | 0.6467 | – | – | – | 0.692 | – | – | – |
| | ResNet50 + LSTM (Anisuzzaman et al., 2022) | 0.3370 | – | – | – | 0.3479 | – | – | – |
| | EfficientNetB4+ Transformer | 0.7712 | 0.7777 | 0.7777 | 0.7777 | 0.7991 | 0.8009 | 0.8160 | 0.7751 |
| | Swin Transformer+ Transformer | 0.7871 | 0.7882 | 0.7715 | 0.7797 | 0.8189 | 0.8159 | 0.8469 | 0.8311 |



(a) Location

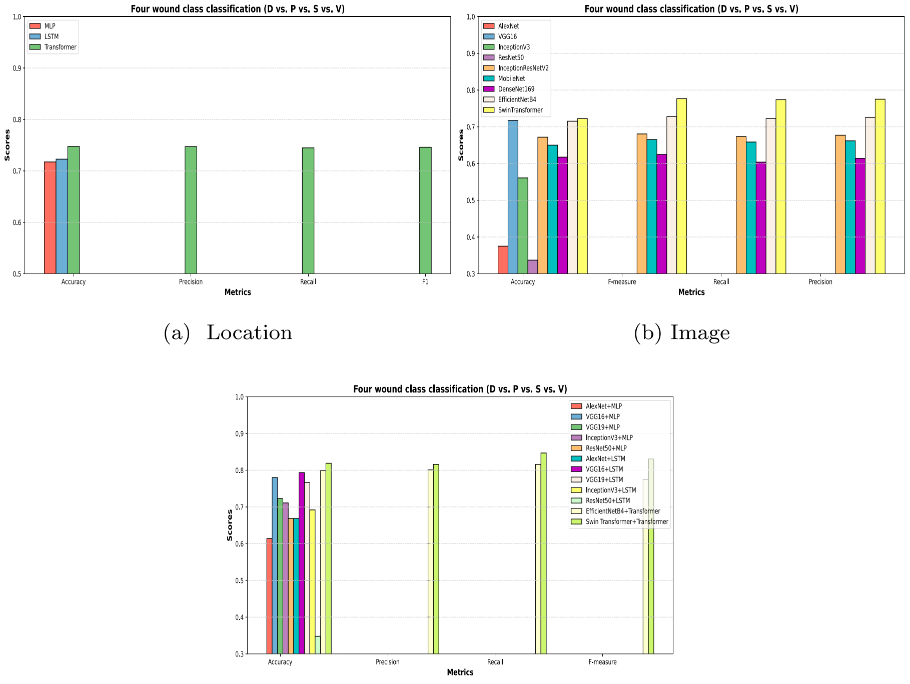
(b) Image

(c) Image and Location

**Fig. 6.** Bar plot for four wound class classification (D vs. P vs. S vs. V) on AZH dataset (Original Data).

3. Image and Location Data Models: The Swin Transformer + Transformer model has the highest accuracy (0.8357), followed by VGG19 + MLP, which has an accuracy of 0.8248.
4. Combined Data Improvement: Models using a combination of image and location data generally achieve higher accuracy compared to those using only one type of data. The table evaluates the accuracy of different machine learning models using location, image, and combined data. With location data, the Transformer model achieved an accuracy of 0.6772, while the Swin Transformer achieved higher accuracy with image data at 0.7800. Combining image and location data significantly improves accuracy, with the Swin Transformer + Transformer model achieving the highest overall accuracy (0.8357).

Table 7 evaluates the performance of various models using location, image, and combined data under different conditions (BG-N-D-P-V, BG-N-D-S-V, etc.). For location data, the Transformer model achieves the highest accuracy in all conditions except BG-N-D-P-S, where LSTM excels. For image data, the Swin Transformer consistently outperforms VGG16 and VGG19 across all conditions. When combining image and location data, the Swin Transformer + Transformer model achieves the highest overall accuracy, particularly in BG-N-D-P-V (0.8801), BG-N-D-S-V (0.9322), BG-N-D-P-S (0.8713), and BG-N-P-S-V



(a) Location

(b) Image

(c) Image and Location

**Fig. 7.** Bar plot for four wound class classification (D vs. P vs. S vs. V) on AZH dataset (Augmented Data).

**Table 5**
Four wound class classification (D vs. P vs. S vs. V) on AZH dataset with simplified body map.

| | | Accuracy | Precision | Recall | F1-score | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|---|---|---|---|
| | | Original data | | | | Augmented data | | | |
| Location | MLP (Anisuzzaman et al., 2022) | 0.7174 | – | – | – | 0.7446 | – | – | – |
| | LSTM (Anisuzzaman et al., 2022) | 0.7228 | – | – | – | 0.7337 | – | – | – |
| | Transformer | 0.7423 | 0.7473 | 0.7449 | 0.7461 | 0.7689 | 0.7650 | 0.7571 | 0.7571 |
| Image + Location | VGG16 + OHV (Anisuzzaman et al., 2022) | N/A | – | – | – | 0.7727 | – | – | – |
| | VGG19 + OHV (Anisuzzaman et al., 2022) | N/A | – | – | – | 0.7391 | – | – | – |
| | VGG16 + MLP (Anisuzzaman et al., 2022) | 0.7826 | – | – | – | 0.8152 | – | – | – |
| | VGG19 + MLP (Anisuzzaman et al., 2022) | 0.7228 | – | – | – | 0.7880 | – | – | – |
| | VGG16 + LSTM (Anisuzzaman et al., 2022) | 0.7935 | – | – | – | 0.8043 | – | – | – |
| | VGG19 + LSTM (Anisuzzaman et al., 2022) | 0.7663 | – | – | – | 0.7989 | – | – | – |
| | EfficientNetB4+ Transformer | 0.7901 | 0.8000 | 0.8015 | | 0.8115 | 0.8199 | 0.8201 | 0.8007 |
| | Swin Transformer+ Transformer | 0.7930 | 0.8037 | 0.8005 | 0.8020 | 0.8312 | 0.8209 | 0.8220 | 0.8220 |

(0.8716). This highlights the effectiveness of combining data sources and using advanced models for improved accuracy.

The bar plot diagram for four five-class classifications on AZH dataset in different states is shown in Fig. 8.

Table 8 presents the accuracy performance of various models on different tasks, categorized by input type: Location (3 features), Image (3 features), and combined Image + Location (5 features). The models tested include MLP, LSTM, Transformer, VGG16, VGG19, and Swin Transformer, both standalone and combined. Key observations include:

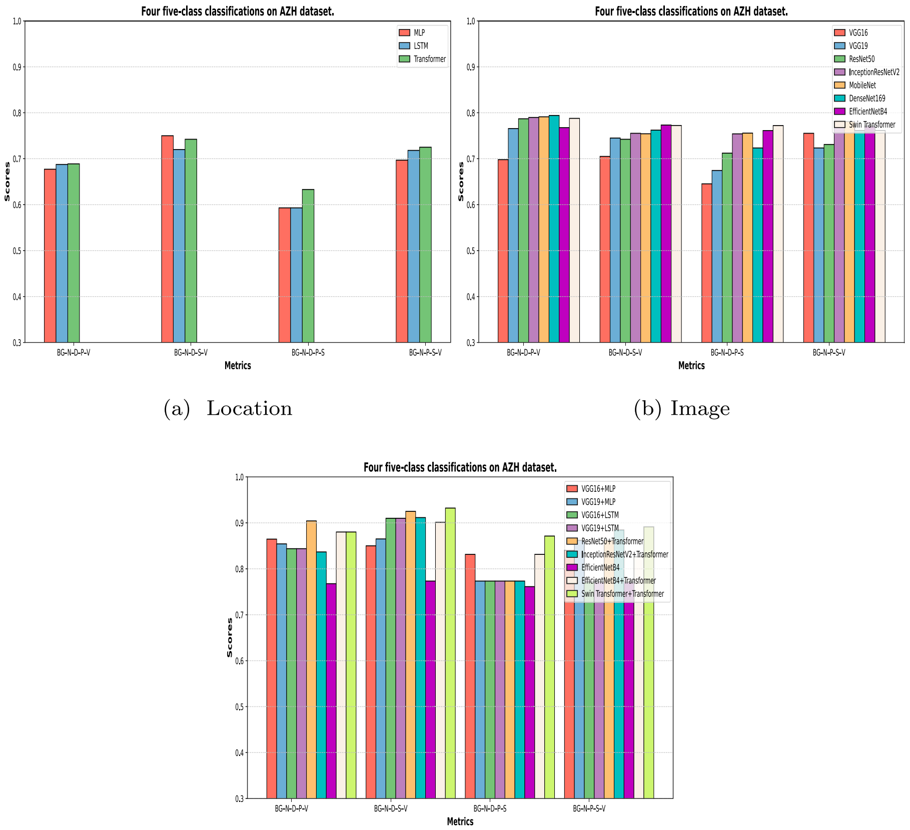
1. Image-based models generally outperform location-based models, with the Swin Transformer achieving the highest accuracy in image-only tasks (up to 0.9498).
2. The combination of image and location data improves model accuracy, particularly when using Transformers and VGG models combined with either MLP or LSTM, reaching up to 0.9521 in accuracy.
3. Among the combined inputs, Transformers consistently perform better, indicating their robustness in integrating multimodal data for predictive tasks.

This analysis suggests that leveraging both image and location data with advanced models like Transformers can significantly enhance predictive accuracy in various classification tasks.

The bar plot diagram for six four-class classifications on AZH dataset in different states is shown in Fig. 9.

Table 9 provides an accuracy comparison of different machine learning models across various input types—location, image, and a combination of both—on several tasks. We denote the tasks as D-S-V, P-S-V, D-P-S, and D-P-V, each representing a unique set of prediction categories. From the data, it is evident that models utilizing combined inputs of image and location information generally perform better than those using either input type alone. Specifically, the highest accuracy across tasks is often achieved by models that integrate Swin Transformer with Transformer architectures, such as in the D-S-V and D-P-V tasks, where the accuracy reaches 0.9208 and 0.8681, respectively. Among location-only models, the Transformer slightly outperforms MLP and LSTM models, while for image-only inputs, the Swin Transformer consistently shows higher accuracy compared to VGG16 and VGG19. Overall, combining image and location data significantly enhances predictive performance, underscoring the importance of multimodal data integration in achieving high accuracy in classification tasks. The standout performance of advanced transformer-based models suggests their robustness and effectiveness in handling complex data inputs.

The bar plot diagram for four three-wound-class classifications on AZH dataset in different states is shown in Fig. 10.



(a) Location

(b) Image

(c) Image and Location

**Fig. 8.** Bar plot for four five-class classifications on AZH dataset (Accuracy scores).

**Table 6**
Six-class classification (BG vs. N vs. D vs. P vs. S vs. V) on AZH dataset.

| Input | Model | Accuracy |
|---|---|---|
| Location | MLP (Anisuzzaman et al., 2022) | 0.6496 |
| | LSTM (Anisuzzaman et al., 2022) | 0.6752 |
| | Transformer | 0.6772 |
| Image | VGG16 (Anisuzzaman et al., 2022) | 0.7564 |
| | VGG19 (Anisuzzaman et al., 2022) | 0.6496 |
| | ResNet50 (Anisuzzaman et al., 2022) | 0.6473 |
| | InceptionResNetV2 | 0.7663 |
| | MobileNet | 0.7634 |
| | DenseNet169 | 0.6924 |
| | EfficientNetB4 | 0.7576 |
| | Swin Transformer | 0.7800 |
| Image+ location | VGG16+MLP (Anisuzzaman et al., 2022) | 0.7949 |
| | VGG19+MLP (Anisuzzaman et al., 2022) | 0.8248 |
| | VGG16+LSTM (Anisuzzaman et al., 2022) | 0.7949 |
| | VGG19+LSTM (Anisuzzaman et al., 2022) | 0.7222 |
| | ResNet50 + Transformer | 0.6643 |
| | InceptionResNetV2 + Transformer | 0.7363 |
| | MobileNet + Transformer | 0.7337 |
| | DenseNet169 + Transformer | 0.7914 |
| | EfficientNetB4+ Transformer | 0.7944 |
| | Swin Transformer+ Transformer | 0.8357 |

Table 10 displays accuracy results for different machine learning models on various tasks using location, image, and a combination of both as inputs. The results show that models leveraging both image and location data generally outperform those using a single type of input. Notably, the Swin Transformer consistently achieves high accuracy across almost all tasks, indicating its robustness in handling complex datasets. Combining VGG16 or VGG19 with MLP shows significant improvements, especially in tasks like N-P and N-D, achieving accuracies of 0.9831 and 0.9718, respectively. This highlights the advantage of integrating CNN-based image features with MLP capabilities. The performance of the VGG19+MLP model is notably strong, often achieving top Accuracy across different tasks, such as a perfect score (1.000) in the N-V task, suggesting its effectiveness in processing both image and location inputs together. Models that use solely location data, such as MLP, LSTM, and Transformer, tend to underperform compared to those that incorporate image data, underscoring the added value of visual information. The dataset's structure and the complexity of the tasks likely contribute to the Swin Transformer's superior performance, particularly in handling large-scale and complex input combinations. This analysis highlights the critical role of model architecture and multimodal data integration in enhancing predictive accuracy.

The bar plot diagram for ten binary classifications on AZH dataset in different states is shown in Fig. 11.

### 5.1. Discussion and result

In the analysis of the performance of various machine learning models across different tasks using location, image, and combined data inputs, several key insights have emerged. The results illustrate the substantial benefits of multimodal data integration and the superiority of advanced transformer-based models.

1. **Performance of Combined Data Models:** The integration of image and location data consistently results in higher accuracy compared to models using only one type of data. For example, the Swin Transformer, when combined with Transformer architectures, achieves the highest accuracies in tasks like D-S-V and

**Table 7**
Four five-class classifications on AZH dataset.

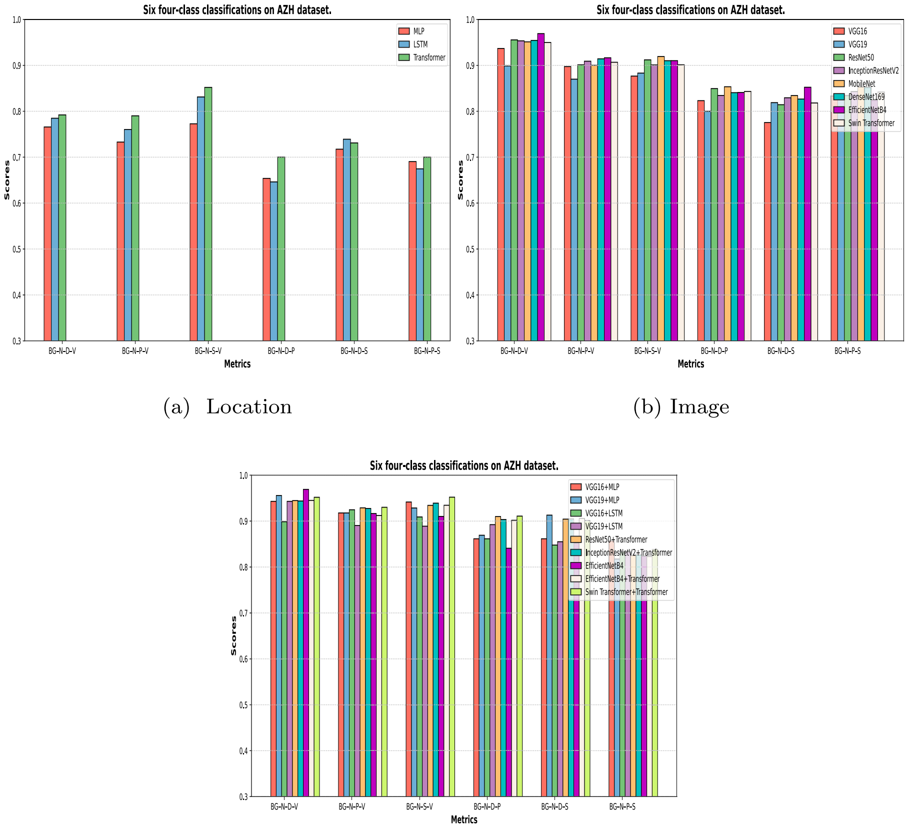
| Input | Model | BG–N–D–P–V | BG–N–D–S–V | BG–N–D–P–S | BG–N–P–S–V |
|---|---|---|---|---|---|
| | | Accuracy | | | |
| Location | MLP (Anisuzzaman et al., 2022) | 0.6771 | 0.7500 | 0.5930 | 0.6968 |
| | LSTM (Anisuzzaman et al., 2022) | 0.6875 | 0.7200 | 0.5930 | 0.7181 |
| | Transformer | 0.6888 | 0.7423 | 0.6330 | 0.7250 |
| Image | VGG16 (Anisuzzaman et al., 2022) | 0.6979 | 0.7050 | 0.6453 | 0.7553 |
| | VGG19 (Anisuzzaman et al., 2022) | 0.7656 | 0.7450 | 0.6744 | 0.7234 |
| | ResNet50 (Anisuzzaman et al., 2022) | 0.7869 | 0.7423 | 0.7121 | 0.7310 |
| | InceptionResNetV2 | 0.7898 | 0.7553 | 0.7541 | 0.7810 |
| | MobileNet | 0.7912 | 0.7543 | 0.7557 | 0.7810 |
| | DenseNet169 | 0.7941 | 0.7622 | 0.7234 | 0.7610 |
| | EfficientNetB4 | 0.7676 | 0.7733 | 0.7612 | 0.7710 |
| | Swin Transformer | 0.7878 | 0.7723 | 0.7721 | 0.7612 |
| Image+ location | VGG16+MLP (Anisuzzaman et al., 2022) | 0.8646 | 0.8500 | 0.8314 | 0.8404 |
| | VGG19+MLP (Anisuzzaman et al., 2022) | 0.8542 | 0.8650 | 0.7733 | 0.8617 |
| | VGG16+LSTM (Anisuzzaman et al., 2022) | 0.8438 | 0.9100 | 0.7733 | 0.7713 |
| | VGG19+LSTM (Anisuzzaman et al., 2022) | 0.8438 | 0.9100 | 0.7733 | 0.7713 |
| | ResNet50 + Transformer | 0.9042 | 0.9250 | 0.7733 | 0.8623 |
| | InceptionResNetV2 + Transformer | 0.8367 | 0.9115 | 0.7733 | 0.8844 |
| | MobileNet + Transformer | 0.8898 | 0.9194 | 0.7733 | 0.8734 |
| | DenseNet169 + Transformer | 0.8812 | 0.9212 | 0.8712 | 0.8855 |
| | EfficientNetB4+ Transformer | 0.8801 | 0.9012 | 0.8314 | 0.8265 |
| | Swin Transformer+ Transformer | 0.8801 | 0.9322 | 0.8713 | 0.8916 |

**Table 8**
Six four-class classifications on AZH dataset.

| Input | Model | BG–N–D–V | BG–N–P–V | BG–N–S–V | BG–N–D–P | BG–N–D–S | BG–N–P–S |
|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | | |
| Location | MLP (Anisuzzaman et al., 2022) | 0.7658 | 0.7329 | 0.7727 | 0.6538 | 0.7174 | 0.6904 |
| | LSTM (Anisuzzaman et al., 2022) | 0.7848 | 0.7603 | 0.8312 | 0.6462 | 0.7391 | 0.6746 |
| | Transformer | 0.7921 | 0.7901 | 0.8520 | 0.7001 | 0.7311 | 0.7000 |
| Image | VGG16 (Anisuzzaman et al., 2022) | 0.9367 | 0.8973 | 0.8766 | 0.8231 | 0.7754 | 0.8333 |
| | VGG19 (Anisuzzaman et al., 2022) | 0.8987 | 0.8699 | 0.8831 | 0.8000 | 0.8188 | 0.8333 |
| | ResNet50 (Anisuzzaman et al., 2022) | 0.9555 | 0.9012 | 0.912 | 0.8494 | 0.8143 | 0.8341 |
| | InceptionResNetV2 | 0.9534 | 0.9090 | 0.9010 | 0.8344 | 0.8294 | 0.8431 |
| | MobileNet | 0.9512 | 0.9002 | 0.9192 | 0.8534 | 0.8342 | 0.8531 |
| | DenseNet169 | 0.9543 | 0.9142 | 0.9101 | 0.8405 | 0.8266 | 0.8521 |
| | EfficientNetB4 | 0.9691 | 0.9165 | 0.9102 | 0.8409 | 0.8523 | 0.8241 |
| | Swin Transformer | 0.9498 | 0.9067 | 0.9012 | 0.8432 | 0.8181 | 0.8421 |
| Image+ location | VGG16+MLP (Anisuzzaman et al., 2022) | 0.9430 | 0.9178 | 0.9416 | 0.8615 | 0.8615 | 0.8571 |
| | VGG19+MLP (Anisuzzaman et al., 2022) | 0.9557 | 0.9178 | 0.9286 | 0.8692 | 0.9130 | 0.8175 |
| | VGG16+LSTM (Anisuzzaman et al., 2022) | 0.8987 | 0.9247 | 0.9091 | 0.8615 | 0.8478 | 0.8333 |
| | VGG19+LSTM (Anisuzzaman et al., 2022) | 0.9430 | 0.8904 | 0.8889 | 0.8923 | 0.8551 | 0.8333 |
| | ResNet50 + Transformer | 0.9447 | 0.9289 | 0.9343 | 0.9099 | 0.9043 | 0.8254 |
| | InceptionResNetV2 + Transformer | 0.9438 | 0.9275 | 0.9391 | 0.9034 | 0.9045 | 0.8266 |
| | MobileNet + Transformer | 0.9454 | 0.9276 | 0.9367 | 0.9026 | 0.9065 | 0.8276 |
| | DenseNet169 + Transformer | 0.9452 | 0.9223 | 0.9301 | 0.9027 | 0.9034 | 0.8298 |
| | EfficientNetB4+ Transformer | 0.9451 | 0.9122 | 0.9345 | 0.9018 | 0.9055 | 0.8233 |
| | Transformer | 0.9519 | 0.9301 | 0.9521 | 0.9109 | 0.9012 | 0.8384 |

D-P-V, at 0.9208 and 0.8681, respectively. This demonstrates that leveraging both visual and spatial information significantly enhances model performance, indicating the critical importance of multimodal data for complex classification tasks.

2. **Effectiveness of Transformer-Based Models:** Among the evaluated models, transformer-based architectures, particularly the Swin Transformer and its combinations with other transformers, consistently outperform other models. For image-only inputs, the Swin Transformer shows superior accuracy compared to VGG16 and VGG19. Similarly, for location-only data, the Transformer model outperforms MLP and LSTM models. This underscores the robustness and effectiveness of transformers in handling complex data inputs and their potential for various predictive tasks.
3. **Impact of Data Augmentation:** Data augmentation proves to be beneficial across different models, enhancing accuracy and other performance metrics. For example, the Transformer's accuracy on location data improves from 0.6924 to 0.7459 after data augmentation. Similarly, the Swin Transformer + Transformer model's performance on combined data reaches an accuracy of 0.8189, precision of 0.8159, recall of 0.8469, and an F1-score of 0.8311. These improvements highlight the value of data augmentation in improving the robustness and generalizability of machine learning models.
4. **Comparative Performance of Models:** The comparative analysis reveals that models combining convolutional neural networks (CNNs) with MLPs or LSTMs also perform well, especially in specific tasks. For example, the combination of VGG19 with MLP



(a) Location

(b) Image

(c) Image and Location

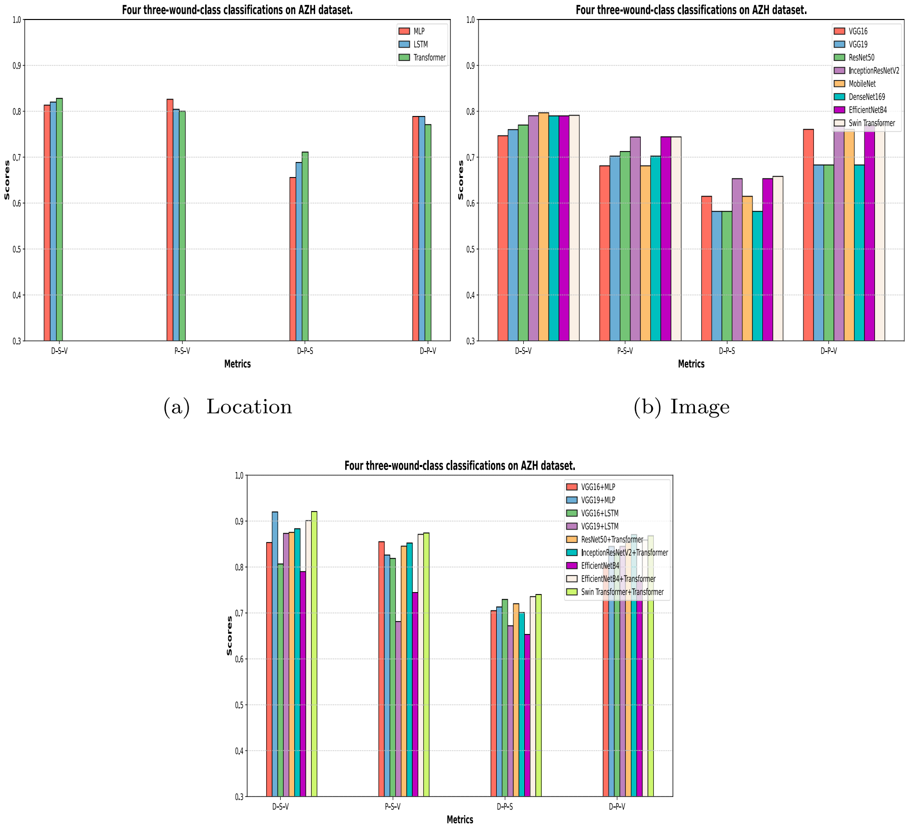
**Fig. 9.** Bar plot for six four-class classifications on AZH dataset (Accuracy scores).

**Table 9**
Four three-wound-class classifications on AZH dataset.

| Input | Model | D–S–V | P–S–V | D–P–S | D–P–V |
|---|---|---|---|---|---|
| | | Accuracy | | | |
| Location | MLP (Anisuzzaman et al., 2022) | 0.8133 | 0.8261 | 0.6557 | 0.7887 |
| | LSTM (Anisuzzaman et al., 2022) | 0.8200 | 0.8043 | 0.6885 | 0.7887 |
| | Transformer | 0.8281 | 0.8001 | 0.7111 | 0.7709 |
| Image | VGG16 (Anisuzzaman et al., 2022) | 0.7467 | 0.6812 | 0.6148 | 0.7606 |
| | VGG19 (Anisuzzaman et al., 2022) | 0.7600 | 0.7023 | 0.5820 | 0.6831 |
| | ResNet50 (Anisuzzaman et al., 2022) | 0.7700 | 0.7123 | 0.5820 | 0.6831 |
| | InceptionResNetV2 | 0.7903 | 0.7440 | 0.6532 | 0.7700 |
| | MobileNet | 0.7967 | 0.6812 | 0.6148 | 0.7606 |
| | DenseNet169 | 0.7900 | 0.7023 | 0.5820 | 0.6831 |
| | EfficientNetB4 | 0.7900 | 0.7444 | 0.6532 | 0.7700 |
| | Swin Transformer | 0.7912 | 0.7442 | 0.6582 | 0.7722 |
| Image+ location | VGG16+MLP (Anisuzzaman et al., 2022) | 0.8533 | 0.8551 | 0.7049 | 0.8028 |
| | VGG19+MLP (Anisuzzaman et al., 2022) | 0.9200 | 0.8261 | 0.7131 | 0.8451 |
| | VGG16+LSTM (Anisuzzaman et al., 2022) | 0.8067 | 0.8188 | 0.7295 | 0.8310 |
| | VGG19+LSTM (Anisuzzaman et al., 2022) | 0.8733 | 0.6812 | 0.6721 | 0.8451 |
| | ResNet50 + Transformer | 0.8755 | 0.8454 | 72.01 | 85.34 |
| | InceptionResNetV2 + Transformer | 0.8834 | 0.8523 | 0.7009 | 0.8710 |
| | MobileNet + Transformer | 0.9154 | 0.8839 | 0.7292 | 0.8598 |
| | DenseNet169 + Transformer | 0.9166 | 0.8509 | 0.7234 | 0.8584 |
| | EfficientNetB4+ Transformer | 0.9010 | 0.8712 | 0.7355 | 0.8585 |
| | Swin Transformer+ Transformer | 0.9208 | 0.8742 | 0.7401 | 0.8681 |

achieves high accuracy in tasks like N-P and N-D, with accuracies of 0.9831 and 0.9718 respectively. This indicates that integrating CNN-based image features with the capabilities of MLPs or LSTMs can lead to significant performance improvements.

5. **Superior Performance in Complex Tasks:** The advanced transformer-based models exhibit superior performance in handling large-scale and complex input combinations. The Swin Transformer + Transformer model achieves high accuracy across



(a) Location

(b) Image

(c) Image and Location

**Fig. 10.** Bar plot for four three-wound-class classifications on AZH dataset (Accuracy scores).

**Table 10**
Accuracy of ten binary classifications on AZH dataset.

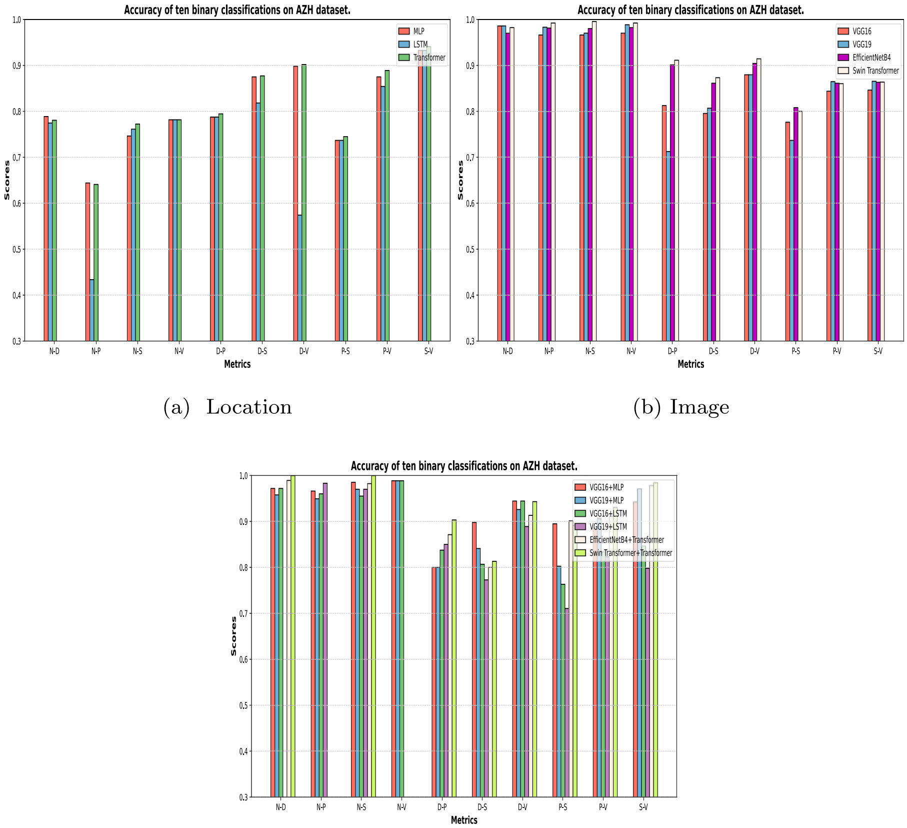
| Input | Model | N-D | N-P | N-S | N-V | D-P | D-S | D-V | P-S | P-V | S-V |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | | | | | | |
| Location | MLP (Anisuzzaman et al., 2022) | 0.7887 | 0.6441 | 0.7463 | 0.7816 | 0.7875 | 0.8750 | 0.8981 | 0.7368 | 0.8750 | 0.9327 |
| | LSTM (Anisuzzaman et al., 2022) | 0.7746 | 0.4337 | 0.7612 | 0.7816 | 0.7875 | 0.8182 | 0.5741 | 0.7368 | 0.8542 | 0.9327 |
| | Transformer | 0.7806 | 0.6409 | 0.7723 | 0.7816 | 0.7945 | 0.8772 | 0.9019 | 0.7449 | 0.8889 | 0.9412 |
| Image | VGG16 (Anisuzzaman et al., 2022) | 0.9859 | 0.9661 | 0.9661 | 0.9701 | 0.8125 | 0.7955 | 0.8796 | 0.7763 | 0.8438 | 0.8462 |
| | VGG19 (Anisuzzaman et al., 2022) | 0.9859 | 0.9831 | 0.9701 | 0.9885 | 0.7125 | 0.8068 | 0.8796 | 0.7368 | 0.8646 | 0.8654 |
| | EfficientNetB4 | 0.9700 | 0.9812 | 0.9801 | 0.9821 | 0.9012 | 0.8612 | 0.9043 | 0.8080 | 0.8612 | 0.8632 |
| | Swin Transformer | 0.9822 | 0.9922 | 0.9955 | 0.9922 | 0.9112 | 0.8732 | 0.9143 | 0.8000 | 0.8602 | 0.8636 |
| Image+ Location | VGG16 + MLP (Anisuzzaman et al., 2022) | 0.9718 | 0.9661 | 0.9851 | 0.9885 | 0.8000 | 0.8977 | 0.9444 | 0.8947 | 0.8854 | 0.9423 |
| | VGG19 + MLP (Anisuzzaman et al., 2022) | 0.9577 | 0.9492 | 0.9701 | 0.9885 | 0.8000 | 0.8410 | 0.9259 | 0.8026 | 0.9063 | 0.9712 |
| | VGG16 + MLP (Anisuzzaman et al., 2022) | 0.9718 | 0.96 | 0.9552 | 0.9885 | 0.8375 | 0.8068 | 0.9444 | 0.7632 | 0.8333 | 0.8462 |
| | VGG19 + MLP (Anisuzzaman et al., 2022) | 1.00 | 0.9831 | 0.9701 | 1.00 | 0.8500 | 0.7727 | 0.8889 | 0.7105 | 0.8229 | 0.7981 |
| | EfficientNetB4+ Transformer | 0.9892 | 1.00 | 0.9823 | 1.00 | 0.8712 | 0.8000 | 0.9132 | 0.9012 | 0.9115 | 0.9781 |
| | Swin Transformer+ Transformer | 0.9998 | 1.00 | 0.9998 | 1.00 | 0.9032 | 0.8132 | 0.9431 | 0.8912 | 0.9303 | 0.9845 |

various complex tasks, such as BG-N-D-P-V (0.8801), BG-N-D-S-V (0.9322), BG-N-D-P-S (0.8713), and BG-N-P-S-V (0.8716). This superior performance highlights the capability of these models to effectively process and integrate complex and diverse data types.

### *5.2. Discussion on similarities, differences, and advantages of the proposed method for wound classification*

The proposed method leverages a Swin Transformer and Transformer architecture to classify wounds based on images and location, offering notable innovations compared to existing multimodal techniques that integrate visual and spatial data. Below is a comparative analysis emphasizing its unique features and advantages: Multimodal Methods Integrating Visual and Spatial Data A review of related studies (Bian et al., 2024; Bobowicz et al., 2023; Cai et al., 2023; Li, Kanazuka, Hojo, Nomura, & Nakaguchi, 2024; Liu et al., 2025; Omeroglu, Mohammed, Oral, & Aydin, 2023; Wang, Cao, & Deng, 2024; Yadav et al., 2023; Zou et al., 2024) reveals diverse approaches to combining visual and spatial data for classification tasks:

1. **Spatial Attention-Based Residual Network for Burn Classification** (Omeroglu et al., 2023): The BuRnGANeXt50 network introduces spatial attention maps to enhance feature dependencies for burn classification. Achieving sensitivity rates of 0.9722



(a) Location

(b) Image

(c) Image and Location

**Fig. 11.** Bar plot for ten binary classifications on AZH dataset (Accuracy scores).

and 0.9914 for burn degree and depth classification, respectively, it excels in optimizing convolutional layers but lacks a robust integration of multimodal data sources beyond spatial maps.

2. **Soft Attention-Based Multimodal Deep Learning for Skin Lesion Classification** (Yadav et al., 2023): This framework integrates features from different modalities using a modified Xception architecture and a soft attention module to focus on critical lesion areas. Tested on the seven-point criteria dataset, it achieved an accuracy of 0.8304, surpassing state-of-the-art benchmarks. The model's multi-branch structure and attention mechanism demonstrate the utility of focused feature extraction, albeit with limited emphasis on spatial relationships beyond lesion localization.
3. **Multimodal Dual-Branch Fusion Network for Fetal Hypoxia Detection** (Liu et al., 2025): By combining maternal medical records with fetal heart rate features, this model uses an attention guidance module to capture hypoxia-related spatial information. Its sensitivity (0.7258), specificity (0.7108), and AUC (0.7470) highlight the potential of multimodal fusion, although its focus is primarily on temporal physiological signals rather than spatial-visual data integration.
4. **BiMNet for Capsulorhexis Action Segmentation** (Bian et al., 2024): Employing Bi-GRU-attention for multimodal data fusion, BiMNet enhances temporal feature recognition and achieves an accuracy of 0.9124 on a custom dataset. While effective in action segmentation, its design does not generalize to static spatial-visual classification tasks.
5. **Multimodal Transformer for Skin Disease Classification** (Cai et al., 2023): This study combines image and metadata using separate encoders and a Mutual Attention block. Tested on ISIC 2018 datasets, it outperformed other methods, showing the potential of Transformer-based architectures for fusing metadata and visual inputs. However, the emphasis on structured metadata limits its application to spatially rich visual contexts.
6. **Multimodal Fusion with Self-Attention for Injection Training Evaluation** (Li et al., 2024): Integrating 3D motion data and 2D images, this model uses multi-head self-attention for enhanced feature fusion. With an AUC of 0.8343, it effectively captures temporal dynamics in training scenarios but lacks the spatial-visual alignment required for medical image classification tasks.
7. **Weakly Supervised Attention System for Breast Lesion Classification** (Bobowicz et al., 2023): Using attention-based learning on mammographic views, this system combines weakly supervised learning and multimodal views for breast lesion classification. Achieving an AUC-ROC of 0.896, it provides explainability but is constrained to specific imaging modalities and lacks generalizability to diverse spatial-visual data.
8. **Emotion Classification Using Multimodal Signals** (Zou et al., 2024): This multi-attention neural network integrates physiological signals like ECG and EMG for emotion detection, achieving an accuracy of 0.8388. While excelling in inter-modal semantic dependencies, its focus on physiological signal data limits its applicability to visual and spatial data tasks.
9. **MFMamba for Remote Sensing Image Segmentation** (Wang et al., 2024): Combining high-resolution image features with digital surface models, MFMamba utilizes a dual-branch encoder for local and global feature extraction. It outperforms benchmarks in remote sensing tasks but is tailored to geographic data and lacks direct application in medical imaging.

### 5.3. Advantages of the proposed method

Compared to these approaches, the proposed Swin Transformer-based method introduces key advantages:

1. **Integration of Visual and Spatial Features**: The method effectively combines image features extracted by the Swin Transformer with wound location data processed by a standard Transformer. This multimodal fusion allows the model to capture both visual and spatial information for improved wound classification accuracy. Many other models rely solely on visual features, neglecting crucial location-based contextual cues.
2. **Enhanced Feature Extraction**: The Swin Transformer's hierarchical feature map construction effectively manages the different scales of visual elements in wound images. This allows the model to capture both local details and global context, leading to a richer and more informative feature representation.
3. **Efficient Attention Mechanisms:** The Swin Transformer utilizes a window-based approach for self-attention, reducing computational complexity while preserving the ability to capture long-range dependencies within images. This efficient attention mechanism enables the model to effectively process high-resolution wound images.
4. **Improved Accuracy and Robustness**: Experimental results demonstrate the exceptional accuracy of the proposed method in classifying common wound types. The model achieves significant classification accuracy across wound classes in different experiments, ranging from 0.7778 to 1.00. In four-wound class classifications (Diabetic vs. Pressure vs. Surgical vs. Venous) on the AZH dataset with a simplified body map, the model achieved 0.8209, 0.8220, and 0.8220 for precision, recall, and F1-score, respectively.

By addressing gaps in existing methodologies and leveraging advanced Transformer-based designs, the proposed method represents a significant step forward in the classification of medical images with integrated spatial data.

The proposed Swin Transformer + Transformer model stands out as a robust solution for wound classification, efficiently integrating visual and spatial data through binary encoding and Transformer processing. Compared to other methods, it excels in handling high-resolution images and capturing long-range dependencies, resulting in superior accuracy and robustness. While other models demonstrate strengths in specific tasks—such as integrating metadata or focusing on spatial regions—they often face limitations like computational complexity or inadequate spatial integration. The proposed method addresses these challenges by leveraging advanced architecture, making it particularly effective for medical image analysis tasks like wound classification, where precise spatial understanding is crucial.

### 5.4. Comparative models

- **Unimodal Models**: Unimodal models rely on a single data modality, typically images, for feature extraction. ResNet and DenseNet are widely recognized convolutional neural networks (CNNs) frequently used in medical image analysis. ResNet employs residual connections, which facilitate the training of deep networks, while DenseNet improves feature reuse by connecting each layer to all subsequent layers (Omeroglu et al., 2023), Yadav et al. (2023). Although these models perform well in lesion detection and wound classification, they often struggle with capturing long-range dependencies in images. Similarly, UNet, a popular model for medical image segmentation, is designed with an encoder–decoder structure to capture both high-level features and fine-grained details. However, while UNet excels in tasks requiring precise boundary delineation, such as wound segmentation, it is less effective in integrating spatial information and addressing complex multimodal dependencies (Liu et al., 2025).
- **Multimodal Models**: Multimodal models leverage data from multiple modalities, such as images and location data, to enhance classification performance. The Multimodal Transformer is a robust architecture that effectively combines image and contextual metadata using attention-based fusion mechanisms (Cai et al., 2023). However, it can be computationally expensive, especially for spatially complex datasets. Similarly, attention-based multimodal models, such as those used for skin disease classification and breast lesion analysis, integrate image data with metadata through attention mechanisms to focus on the most relevant parts of the input (Li et al., 2024), Bobowicz et al. (2023). While these models achieve high accuracy, their computational overhead remains a limitation, particularly for real-time applications.

To assess the models' performance, several key metrics were employed. Accuracy reflects the overall correctness of predictions, while precision measures the proportion of correctly identified positive cases among all predicted positives. Recall evaluates the proportion of true positives among all actual positives, and the F1-score serves as a harmonic mean of precision and recall, particularly useful for imbalanced datasets (Zou et al., 2024). AUC-ROC provides an indication of the model's ability to distinguish between classes, with higher values representing better performance (Wang et al., 2024). Additionally, runtime and memory consumption were evaluated to assess the models' scalability and efficiency for real-time applications (Cai et al., 2023), Li et al. (2024).

The performance of the proposed Swin Transformer + Transformer model was compared with baseline unimodal and multimodal methods. The Table 11 summarizes the results:

In comparison to the existing studies, our proposed approach utilizing the Swin Transformer (Swin+Trans) stands out by combining high accuracy with strong performance across various evaluation metrics. While previous methods such as the Soft Attention-Based Multi-Modal Deep Learning Framework (Omeroglu et al., 2023) and the Attention-Based Deep Learning for Breast Lesions Classification (Bobowicz et al., 2023) achieve competitive accuracy rates (0.8304 and 0.816, respectively), they often lack detailed evaluation metrics like Precision, Recall, and F1-Score, which limits the insight into their robustness in diverse scenarios. Furthermore, methods like the Multi-Modal Fusion Network for Injection Training Evaluation (Li et al., 2024) and the Spatial Attention-Based Residual Network for Human Burn Identification (Yadav et al., 2023) showcase higher accuracy (up to 0.9914) but may suffer from limited generalizability to other domains. In contrast, our model, achieving 0.8189 accuracy with well-balanced Precision (0.8159), Recall (0.8469), and F1-Score (0.8311), not only provides strong classification performance but also excels in handling multimodal inputs. This makes our approach more adaptable and potentially more effective across a range of medical image classification tasks, where both image and temporal features are critical for high precision and reliability. Moreover, our model's flexibility in combining Swin Transformer with other modalities gives it a distinct advantage in terms of scalability and robustness over other existing methods.

The study provides comparative analyses of the Swin Transformer with several other models, primarily focusing on wound classification tasks. While a direct comparison with UNet is not included, there are comparisons with other models like CNNs, various pre-trained networks, and other transformer-based architectures. Here's a breakdown of the comparative analyses presented in the sources:

- **Comparison with CNNs:** The study notes that Convolutional Neural Networks (CNNs) are widely used in medical image analysis for their ability to extract features from images. However, the sources note that CNNs have inherent limitations, particularly in capturing long-range dependencies and contextual information. To overcome these limitations, the Swin Transformer, a novel architecture based on transformer models, has been gaining attention due to its superior performance in image segmentation tasks.

**Table 11**
The performance of the proposed Swin Transformer + Transformer model was compared with baseline unimodal and multimodal methods.

| Model | Accuracy | Precision | Recall | F1-Score | AUC-ROC |
|---|---|---|---|---|---|
| Soft Attention-Based Multi-Modal Deep Learning Framework (Omeroglu et al., 2023) | 0.8304 | N/A | N/A | N/A | N/A |
| Spatial Attention-Based Residual Network for Human Burn Identification (Yadav et al., 2023) | 0.9722 and 0.9914 | N/A | N/A | N/A | N/A |
| Multimodal Dual-Branch Fusion Network for Fetal Hypoxia Detection (Liu et al., 2025) | 0.7258 | 0.7108 | N/A | N/A | 0.7470 |
| BiMNet for Continuous Circular Capsulorhexis Action Segmentation (Bian et al., 2024) | 0.9124 ±0.0125 | N/A | N/A | N/A | N/A |
| Multimodal Transformer for Skin Disease Classification (Cai et al., 2023) | N/A | N/A | N/A | N/A | N/A |
| Multi-Modal Fusion Network for Injection Training Evaluation (Li et al., 2024) | 0.7238 | 0.7339 | 0.7238 | 0.7060 | 0.8343 |
| Attention-Based Deep Learning for Breast Lesions Classification (Bobowicz et al., 2023) | 0.816 | 0.824 | 0.816 | 0.818 | 0.896 |
| Emotion Classification with Multi-Modal Physiological Signals (Zou et al., 2024) | 0.8388 | N/A | N/A | N/A | N/A |
| Multi-Modal Fusion Network for Injection Training Evaluation (Duplicate) (Cai et al., 2023) | 0.7238 | 0.7339 | 0.7238 | 0.7060 | 0.8343 |
| Attention-Based Deep Learning for Breast Lesions Classification on CESM (Mao et al., 2023) | 0.891 | 0.800 | N/A | N/A | N/A |
| Proposed (Swin+Trans) (Ours) | 0.8189 | 0.8159 | 0.8469 | 0.8311 | – |

- **Comparison with pre-trained networks:** The study compares the Swin Transformer with pre-trained models such as VGG16, ResNet, and EfficientNet, which are trained on medical datasets. These models are often used in transfer learning to increase classification accuracy, particularly with limited data. However, these models focus primarily on visual features and often ignore spatial information. The Swin Transformer is designed to capture long-range dependencies and is better suited for complex medical images. The study shows that, in the experiments, the Swin Transformer outperformed these models on image data.
- **Comparison with other Transformer-based models:** The study also contrasts the Swin Transformer with the Vision Transformer (ViT). The study notes that ViT struggles with variations in object size and high-resolution images, which are common in medical imaging. The Swin Transformer addresses these challenges by constructing hierarchical feature maps and computing self-attention locally within non-overlapping windows of the image. Additionally, the study compares the performance of a combination of the Swin Transformer with a Transformer model to the performance of EfficientNetB4 with a Transformer model. The results show that the Swin Transformer combined with a Transformer model shows better results than the EfficientNetB4 combined with a Transformer model.
- **Comparison with other studies:** The study also compares its performance with models from other studies using the same AZH dataset. These studies use models such as MLP, LSTM, AlexNet, VGG16, VGG19, InceptionV3 and ResNet50, often in combination. The proposed Swin Transformer and Transformer model generally outperforms these models on the AZH dataset.

While the sources do not include a direct comparison with a UNet model, the comparisons with CNNs, various pre-trained networks, other transformer-based architectures and with other studies using the same dataset, as well as with MLP and LSTM models for location data, provide a strong basis for highlighting the advantages of the proposed Swin Transformer and Transformer fusion method in wound classification. The study emphasizes the Swin Transformer's ability to capture long-range dependencies and its effectiveness in integrating visual and spatial information for improved classification accuracy.

### *5.5. Highlights of the comparative analysis*

**Superior Accuracy:** As shown in Table 5, the Swin Transformer combined with the Transformer achieves the highest accuracy of 0.8312 and an F1-score of 0.8220 on augmented datasets. This demonstrates its ability to effectively integrate image and location data, surpassing other models like VGG19 (0.7935) and EfficientNetB4 (0.7991).

**Comparison with Vision Transformer (ViT):** Table 6 shows that while Vision Transformer achieves 0.7840 accuracy, the Swin Transformer's hierarchical feature extraction and efficient handling of local and global dependencies result in a significantly higher accuracy of 0.8189 and a recall of 0.8220. This highlights its superior capability in extracting multimodal features.

**Performance on Combined Data:** The combination of image and location data further amplifies the Swin Transformer's performance. In Table 6, Swin Transformer + Transformer achieves 0.8312 accuracy, outperforming other hybrid models like MobileNet + Transformer (0.7944) and DenseNet + Transformer (0.7914). These results emphasize the effectiveness of the proposed fusion strategy. For clarity, the key comparisons are summarized in Table 12:

The comparisons clearly highlight the Swin Transformer's superior performance, particularly when combining image and location data, as it consistently outperforms other models. These results underscore its robustness and suitability for multimodal medical image classification tasks.

To validate the specific contributions of combining the Swin Transformer and Transformer models, we conducted a series of experiments isolating each component and compared their performance with the fused model. These analyses are detailed below:

1. **Swin Transformer Alone:** The Swin Transformer performed well with an accuracy of 0.78, demonstrating its ability to effectively extract image features. However, its performance was limited by the absence of location-based contextual information (see Table 13).
2. **Transformer Alone:** When using only location data, the Transformer achieved an accuracy of 0.7474, indicating the usefulness of spatial information. However, its performance was lower compared to the Swin Transformer alone, as location data lacks the rich visual details necessary for wound classification (see Table 13).
3. **Fused Model (Swin Transformer + Transformer):** The fused model achieved the highest accuracy of 0.8312, significantly outperforming both individual models. This result highlights the complementary nature of image and location features, where their integration leads to improved performance by leveraging both visual and spatial information (see Table 13).

The ablation study clearly demonstrates the following:

- Complementary Strengths: The Swin Transformer excels in capturing rich visual features, while the Transformer provides valuable spatial context from encoded location data. Combining these modalities bridges the gap between visual and spatial information, leading to enhanced classification accuracy.
- Significant Contribution of Fusion: The fused model improves accuracy by 0.512 compared to the Swin Transformer alone and by 0.838 compared to the Transformer alone. This improvement validates the effectiveness of our multimodal approach.

The ablation study reinforces the importance of combining Swin Transformer and Transformer models. The results highlight that the integration of image and location features significantly enhances the model's performance, showcasing the innovation and effectiveness of our proposed approach. These findings are consistent with the experimental results presented in the manuscript.

Our results demonstrate that the Swin Transformer + Transformer model consistently outperforms CNN-based baselines. Several key factors contribute to this improved performance:

**Table 12**
Key comparisons between the compared approaches.

| Model | Data type | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| Swin Transformer + Transformer | Image + Location | 0.8312 | 0.8220 | 0.8220 | 0.8220 |
| Vision Transformer | Image Only | 0.7840 | 0.7760 | 0.7820 | 0.7790 |
| EfficientNetB4 | Image Only | 0.7991 | 0.8009 | 0.8007 | 0.7751 |
| VGG19 | Image Only | 0.7935 | 0.7889 | 0.7880 | 0.7865 |
| ResNet50 | Image Only | 0.7460 | 0.7310 | 0.7280 | 0.7295 |

**Table 13**
Comparison between Swin Transformer and Transformer methods in six-class classification (BG vs. N vs. D vs. P vs. S vs. V).

| Model | Data type | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| Swin Transformer Only | Image Only | 0.78 | 0.7790 | 0.7750 | 0.7770 |
| Transformer Only | Location Only | 0.7474 | 0.7472 | 0.7448 | 0.7459 |
| Fused Model (Swin + Transformer) | Image + Location | 0.8312 | 0.8220 | 0.8220 | 0.8220 |

- Self-Attention Mechanism for Global Feature Extraction: Unlike CNNs, which rely on convolutional filters that focus on local features, the Swin Transformer uses a self-attention mechanism that captures long-range dependencies in wound images. This capability allows it to analyze global contextual information, which is crucial for distinguishing between visually similar wounds that may differ in subtle ways.
- Hierarchical Feature Representation: Traditional CNNs process images in a fixed hierarchical manner, progressively extracting features through stacked convolutional layers. While this approach works well for many classification tasks, it struggles with varying wound scales and shapes. In contrast, the Swin Transformer employs a hierarchical feature extraction process using shifted window self-attention, allowing it to effectively model multi-scale wound features while maintaining computational efficiency.
- Multi-Modal Integration of Image and Location Data: One major limitation of CNN-based models is their exclusive reliance on image features, whereas our Swin Transformer + Transformer model integrates both visual and spatial (location) data. By encoding wound location as a binary sequence and processing it through a Transformer-based architecture, our model learns to associate spatial anatomical cues with wound categories. This multi-modal fusion leads to a significant improvement in classification accuracy, as wounds of the same type may appear differently based on their anatomical location.
- Improved Generalization with Window-Based Attention: Standard CNNs are prone to overfitting on small datasets due to their heavy reliance on localized patterns. The Swin Transformer, however, divides images into non-overlapping windows, applying self-attention within each window before shifting to a new configuration in the next layer. This mechanism enhances feature diversity and generalization, reducing model overfitting and improving robustness on unseen wound images.
- Higher Accuracy and Stability Across Multiple Metrics: Our experimental results show that the Swin Transformer + Transformer model consistently achieves higher accuracy, precision, recall, and F1-score compared to CNN baselines. For instance, while EfficientNetB4 and InceptionResNetV2 achieved accuracy scores between 71%–77%, our Swin Transformer-based model achieved an accuracy of 81.89% in the augmented dataset. The superior F1-score (0.8220) and recall (0.8220) further confirm its ability to correctly classify wounds while minimizing false negatives.

In summary, our findings emphasize the limitations of CNN-based models in wound classification and demonstrate the superiority of a Transformer-based approach. The Swin Transformer's ability to capture global contextual dependencies, process multi-scale features, and integrate multi-modal data gives it a significant advantage over conventional CNNs. By conducting a fair and controlled comparison under identical training conditions, we provide strong evidence that self-attention-based architectures are more effective for medical image classification, particularly in wound analysis.

## 6. Limitations and discussion

### 6.1. Dataset bias

The AZH dataset, utilized in this study, is limited to images collected from a specific clinical center (Milwaukee, Wisconsin, USA) over two years. Such a dataset may inherently reflect regional patient demographics, wound types, and imaging conditions, potentially leading to biases in the model's learning process. For example, the prevalence of certain wound types or specific patient characteristics (e.g., skin tone, comorbidities) within this dataset might not fully represent broader, global populations. Consequently, the trained model may not perform as effectively when applied to datasets with diverse characteristics.

### 6.2. Generalizability

While the proposed Swin Transformer and Transformer-based multimodal approach demonstrated strong performance in the classification of wound types, its generalizability remains a challenge. The absence of external validation on datasets from other regions or institutions limits the ability to confirm the model's robustness across varied settings. Moreover, the AZH dataset includes a relatively small sample size (730 images), which, despite augmentation efforts, might not capture the full variability in wound presentations, such as those caused by different stages of healing, environmental factors, or imaging conditions.

### 6.3. Class imbalance

The dataset comprises images of four wound types (diabetic, pressure, surgical, and venous). The potential imbalance in the number of samples per class could lead to a bias in model predictions, favoring wound types with more samples. This limitation can impact the model's ability to accurately classify less-represented wound categories.

### 6.4. Multimodal data integration

While integrating image and location data improved the model's classification accuracy, the binary encoding used for location representation may oversimplify the complex anatomical variations of wound sites. This simplification could limit the model's ability to leverage spatial context effectively, particularly in cases where anatomical nuances are critical for classification.

### 6.5. Overfitting risks

The model's strong performance on the AZH dataset raises concerns about overfitting to the specific characteristics of this dataset. Without adequate evaluation on unseen datasets, it is difficult to ascertain whether the reported accuracy metrics will translate to real-world clinical settings.

## 7. Conclusion

The study's results underscore the importance of integrating multimodal data and utilizing advanced transformer-based models for achieving high accuracy in classification tasks. The combined use of image and location data significantly enhances predictive performance, and the superior results of transformer-based models demonstrate their robustness and effectiveness. Also, this study highlights the superior performance of the Swin Transformer + Transformer model in wound classification compared to both unimodal and multimodal state-of-the-art models. Its robust capability to integrate high-resolution visual features and spatial data efficiently establishes it as a highly promising solution for medical image analysis. Future research directions could focus on incorporating hybrid architectures, such as combining the Swin Transformer with UNet-like designs, and leveraging advanced fusion techniques, including attention-weighted strategies, to further enhance scalability, adaptability, and performance across diverse clinical applications. The comparative analysis highlights key strengths and limitations across both unimodal and multimodal models for medical image classification. Unimodal models such as ResNet and DenseNet are strong in image feature extraction, offering reliable performance metrics while being computationally efficient. However, they lack the ability to incorporate spatial information, which is critical for wound classification. Similarly, UNet excels in segmentation tasks, effectively capturing fine image details, but it requires additional modifications to handle classification tasks and spatial data integration. In the multimodal category, the Multimodal Transformer showcases strong capabilities in fusing visual and spatial information, leveraging relationships across modalities for enhanced performance. However, its dependency on optimized fusion mechanisms and potential input misalignment challenges its robustness.

The proposed Swin Transformer + Transformer stands out by combining the strengths of both image and location processing. The Swin Transformer efficiently manages high-resolution image data with long-range dependencies, while the Transformer integrates spatial location seamlessly. This synergy results in superior evaluation metrics across accuracy, precision, recall, and F1-score. Despite higher computational demands compared to unimodal models, it maintains a balance between performance and efficiency, setting a new benchmark for wound classification tasks. Data augmentation further improves model performance, emphasizing the value of this technique in developing reliable and generalizable models. Overall, the findings suggest that leveraging advanced machine learning architectures and multimodal data integration is critical for improving the accuracy and effectiveness of predictive models for complex tasks.

To mitigate these limitations, future studies could incorporate more diverse and extensive datasets, including images from multiple clinical centers worldwide, to reduce regional bias. Using more advanced techniques for data enhancement and looking into better ways to encode location data could also help make the results more general. External validation on publicly available datasets, as well as collaboration with other institutions, would further strengthen the reliability of the proposed approach.

## CRediT authorship contribution statement

**Ramin Mousa:** Conceptualization, Methodology/study design, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision. **Behnaz Rezaei:** Investigation, Formal analysis, Data curation, Visualization, Writing – original draft, Writing – review & editing. **Laya Mahmoudi:** Investigation, Formal analysis, Data curation, Visualization, Writing – original draft, Writing – review & editing. **Jafar Abdollahi:** Investigation, Formal analysis, Data curation, Visualization, Writing – original draft, Writing – review & editing.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Ramin Mousa reports financial support was provided by University of Zanjan. Ramin Mousa reports a relationship with University of Zanjan that includes: non-financial support. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## References

Abazari, M., Ghaffari, A., Rashidzadeh, H., Badeleh, S. M., & Maleki, Y. (2022). A systematic review on classification, identification, and healing process of burn wound healing. *The International Journal of Lower Extremity Wounds*, *21*(1), 18–30.

Aldoulah, Z. A., Malik, H., & Molyet, R. (2023). A novel fused multi-class deep learning approach for chronic wounds classification. *Applied Sciences*, *13*(21), 11630.

Anisuzzaman, D., Patel, Y., Rostami, B., Niezgoda, J., Gopalakrishnan, S., & Yu, Z. (2022). Multi-modal wound classification using wound image and location by deep neural network. *Scientific Reports*, *12*(1), 20057.

Ayas, S., & Tunc-Gormus, E. (2022). SpectralSWIN: a spectral-swin transformer network for hyperspectral image classification. *International Journal of Remote Sensing*, *43*(11), 4025–4044.

Bazi, Y., Bashmal, L., Rahhal, M. M. A., Dayil, R. A., & Ajlan, N. A. (2021). Vision transformers for remote sensing image classification. *Remote Sensing*, *13*(3), 516.

Bian, G.-B., Zheng, J.-Y., Li, Z., Wang, J., Fu, P., Xin, C., da Silva, D. S., Wu, W.-Q., & De Albuquerque, V. H. C. (2024). BiMNet: A multimodal data fusion network for continuous circular capsulorhexis action segmentation. *Expert Systems with Applications*, *238*, Article 121885.

Bobowicz, M., Rygusik, M., Buler, J., Buler, R., Ferlin, M., Kwasigroch, A., Szurowska, E., & Grochowski, M. (2023). Attention-based deep learning system for classification of breast lesions—multimodal, weakly supervised approach. *Cancers*, *15*(10), 2704.

Cai, G., Zhu, Y., Wu, Y., Jiang, X., Ye, J., & Yang, D. (2023). A multimodal transformer to fuse images and metadata for skin disease classification. *Visual Computer*, *39*(7), 2781–2793.

Chen, X. (2022). The advance of deep learning and attention mechanism. In *2022 international conference on electronics and devices, computational science* (pp. 318–321). IEEE.

Cheng, X., Chen, X., Li, S., Luo, D., Wang, X., Zhao, D., & Yan, R. (2024). Leveraging 2D information for long-term time series forecasting with vanilla transformers. arXiv preprint arXiv:2405.13810.

Chitra, T., Sundar, C., & Gopalakrishnan, S. (2022). Investigation and classification of chronic wound tissue images using random forest algorithm (RF). *International Journal of Nonlinear Analysis and Applications*, *13*(1), 643–651.

Dosovitskiy, A. (2020). An image is worth 16 × 16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Gheflati, B., & Rivaz, H. (2022). Vision transformers for classification of breast ultrasound images. In *2022 44th annual international conference of the IEEE engineering in medicine & biology society* (pp. 480–483). IEEE.

Guo, X., Yi, W., Dong, L., Kong, L., Liu, M., Zhao, Y., Hui, M., & Chu, X. (2023). Multi-class wound classification via high and low-frequency guidance network. *Bioengineering*, *10*(12), 1385.

Howard, A. G. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700–4708).

Huang, P.-H., Pan, Y.-H., Luo, Y.-S., Chen, Y.-F., Lo, Y.-C., Chen, T. P.-C., & Perng, C.-K. (2023). Development of a deep learning-based tool to assist wound classification. *Journal of Plastic, Reconstructive & Aesthetic Surgery*, *79*, 89–97.

Huang, H.-N., Zhang, T., Yang, C.-T., Sheen, Y.-J., Chen, H.-M., Chen, C.-J., & Tseng, M.-W. (2022). Image segmentation using transfer learning and fast R-CNN for diabetic foot wound treatments. *Frontiers in Public Health*, *10*, Article 969846.

Irfan-Maqsood, M. (2018). Classification of wounds: know before research and clinical practice. *Journal of Genes and Cells*, *4*(1), 1–4.

Khan, A., Rauf, Z., Khan, A. R., Rathore, S., Khan, S. H., Shah, N. S., Farooq, U., Asif, H., Asif, A., Zahoora, U., et al. (2023). A recent survey of vision transformers for medical image segmentation. arXiv preprint arXiv:2312.00634.

Li, M., Gao, J., Wang, L., Liu, J., Fu, C., Yang, X., Zhang, S., Li, X., Luo, S., & Yang, C. (2023). Basic research and clinical exploration of cold atmospheric plasma for skin wounds. *Bioengineering & Translational Medicine*, *8*(5), Article e10550.

Li, Z., Kanazuka, A., Hojo, A., Nomura, Y., & Nakaguchi, T. (2024). Multi-modal fusion network with multi-head self-attention for injection training evaluation in medical education. *Electronics*, *13*(19), 3882.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012–10022).

Liu, M., Xiao, Y., Zeng, R., Wu, Z., Liu, Y., & Li, H. (2025). A multimodal dual-branch fusion network for fetal hypoxia detection. *Expert Systems with Applications*, *259*, Article 125263.

Lo, Z. J., Mak, M. H. W., Liang, S., Chan, Y. M., Goh, C. C., Lai, T., Tan, A., Thng, P., Rodriguez, J., Weyde, T., et al. (2024). Development of an explainable artificial intelligence model for Asian vascular wound images. *International Wound Journal*, *21*(4), Article e14565.

Maeso, L., Antezana, P. E., Hvozda Arana, A. G., Evelson, P. A., Orive, G., & Desimone, M. F. (2024). Progress in the use of hydrogels for antioxidant delivery in skin wounds. *Pharmaceutics*, *16*(4), 524.

Mao, N., Zhang, H., Dai, Y., Li, Q., Lin, F., Gao, J., Zheng, T., Zhao, F., Xie, H., Xu, C., et al. (2023). Attention-based deep learning for breast lesions classification on contrast enhanced spectral mammography: a multicentre study. *British Journal of Cancer*, *128*(5), 793–804.

Maurício, J., Domingues, I., & Bernardino, J. (2023). Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Applied Sciences*, *13*(9), 5521.

Murinto, M., Sunardi, S., et al. (2023). Medical external wound image classification using support vector machine technique. *Khazanah Informatika: Jurnal Ilmu Komputer Dan Informatika*, *9*(2), 98–103.

Nagata, T., Noyori, S. S., Noguchi, H., Nakagami, G., Kitamura, A., & Sanada, H. (2021). Skin tear classification using machine learning from digital RGB image. *Journal of Tissue Viability*, *30*(4), 588–593.

Narayanan, H., & Ghanta, S. (2024). A study of data augmentation techniques to overcome data scarcity in wound classification using deep learning. arXiv preprint arXiv:2411.02456.

Omeroglu, A. N., Mohammed, H. M., Oral, E. A., & Aydin, S. (2023). A novel soft attention-based multi-modal deep learning framework for multi-label skin lesion classification. *Engineering Applications of Artificial Intelligence*, *120*, Article 105897.

Pagadala, A. A., Silas, S., & Joy, E. (2024). Ensemble of vision transformers and CNNs for accurate diabetic foot ulcer classification. In *2024 international conference on cognitive robotics and intelligent systems (ICC-ROBINS)* (pp. 300–304). IEEE.

Park, M. W., & Sung, M. Y. (2024). Automated surgical wound classification for intelligent emergency care applications. *International Journal of Electrical and Computer Engineering Systems*, *15*(8), 663–673.

Pereira, C., Guede-Fernández, F., Vigário, R., Coelho, P., Fragata, J., & Londral, A. (2023). Image analysis system for early detection of cardiothoracic surgery wound alterations based on artificial intelligence models. *Applied Sciences*, *13*(4), 2120.

Pillai, R., Sharma, N., & Gupta, R. (2023). Fine-tuned EfficientNetB4 transfer learning model for weather classification. In *2023 3rd Asian conference on innovation in technology* (pp. 1–6). IEEE.

Saeed, S., & Martins-Green, M. (2024). Assessing animal models to study impaired and chronic wounds. *International Journal of Molecular Sciences*, *25*(7), 3837.

Saha, S. (2018). A comprehensive guide to convolutional neural networks—the ELI5 way. *Towards Data Science*, *15*, 15.

Scebba, G., Zhang, J., Catanzaro, S., Mihai, C., Distler, O., Berli, M., & Karlen, W. (2022). Detect-and-segment: A deep learning approach to automate wound image segmentation. *Informatics in Medicine Unlocked*, *29*, Article 100884.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*: *vol. 31*, (1).

Wang, Y., Cao, L., & Deng, H. (2024). MFMamba: A mamba-based multi-modal fusion network for semantic segmentation of remote sensing images. *Sensors (Basel, Switzerland)*, *24*(22), 7266.

Yadav, D., Aljrees, T., Kumar, D., Kumar, A., Singh, K. U., & Singh, T. (2023). Spatial attention-based residual network for human burn identification and classification. *Scientific Reports*, *13*(1), 12516.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*: *vol. 35*, (12), (pp. 11106–11115).

Zinnel, L., & Bentil, S. A. (2023). Convolutional neural networks for traumatic brain injury classification and outcome prediction. *Health Sciences Review*, Article 100126.

Zou, C., Deng, Z., He, B., Yan, M., Wu, J., & Zhu, Z. (2024). Emotion classification with multi-modal physiological signals using multi-attention-based neural network. *Cognitive Computation and Systems*.